Titre: Title:	Architecture d'interopérabilité et mécanismes de relève pour les réseaux sans fil de prochaine génération
Auteur: Author:	Christian Wilfrid Makaya
Date:	2007
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Makaya, C. W. (2007). Architecture d'interopérabilité et mécanismes de relève pour les réseaux sans fil de prochaine génération [Ph.D. thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/7995/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/7995/
Directeurs de recherche: Advisors:	
Programme: Program:	Unspecified

UNIVERSITÉ DE MONTRÉAL

ARCHITECTURE D'INTEROPÉRABILITÉ ET MÉCANISMES DE RELÈVE POUR LES RÉSEAUX SANS FIL DE PROCHAINE GÉNÉRATION

CHRISTIAN WILFRID MAKAYA
DÉPARTEMENT DE GÉNIE INFORMATIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE INFORMATIQUE)
SEPTEMBRE 2007

Your file Votre référence
ISBN: 978-0-494-37130-5
Our file Notre référence
ISBN: 978-0-494-37130-5

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée:

ARCHITECTURE D'INTEROPÉRABILITÉ ET MÉCANISMES DE RELÈVE POUR LES RÉSEAUX SANS FIL DE PROCHAINE GÉNÉRATION

présentée par: MAKAYA Christian Wilfrid
en vue de l'obtention du diplôme de: Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de:

M. DAGENAIS Michel, Ph.D., président
M. PIERRE Samuel, Ph.D., membre et directeur de recherche
M. QUINTERO Alejandro, Doct., membre
M. KARMOUCH Ahmed, Ph.D., membre

À la mémoire de ma mère, 20 ans déjà !

À mon père, mes frères et sœurs

REMERCIEMENTS

Tout au long de ce travail, plusieurs personnes m'ont apporté leur aide, soutien et appui. Je prendrai donc cet espace pour les remercier de leur précieuse attention.

Je remercie mon directeur de recherche, le Professeur Samuel Pierre pour son encadrement, ses commentaires et sa disponibilité. Je lui exprime aussi ma reconnaissance pour m'avoir fourni le soutien financier durant ce projet. J'espère du fond du cœur que cette thèse n'est que le début d'une étroite collaboration pour les années à avenir.

Mes remerciements vont aussi aux Professeurs Alejandro Quintero, Ahmed Karmouch et Michel Dagenais d'avoir accepté de participer à mon jury. Malgré leurs multiples occupations, ils m'ont accordé leur temps pour évaluer cette thèse.

Je remercie Yves Lemieux pour ses conseils et les excellentes conditions de travail auprès de Ericsson Research Canada lors de la réalisation de cette thèse.

Mes remerciements s'adressent aussi à tous ceux et toutes celles qui sont loin ou non de l'informatique et des télécommunications qui m'ont soutenu et encouragé. Je ne peux pas tous les citer, la liste étant exhaustive, j'espère qu'ils/elles se reconnaîtront. Un merci particulier à mes amis Christiane Dzongang et Fidèle Likibi.

J'exprime ma profonde et sincère gratitude à mon père, mes frères et sœurs (Alexandre, Nathalie, Davy et Ginette). Le soutien et l'amour indéfectible dont je bénéficie de vous ne peuvent se résumer par des simples mots. Merci pour tout.

Pour terminer, je remercie ma tendre et douce moitié, Anne-Dominique Makosso. Merci pour ton support, ta compréhension et ta patience face à la distance. Merci pour tes mots doux, ton attention, tes encouragements et surtout ton amour.

RÉSUMÉ

La coexistence de plusieurs réseaux de communication et technologies suscite l'intérêt de leur intégration et interopérabilité afin de tirer profit des avantages de chacun d'eux. Parmi les différentes couches de la pile de protocole TCP/IP, la technologie IP (*Internet Protocol*) semble la plus appropriée pour réaliser cette intégration. En effet, elle est la plus basse couche commune aux différentes technologies des réseaux d'accès. Cette intégration sera le fondement de la conception des réseaux de communication dits de prochaine génération, en particulier les réseaux sans fil et mobile. Ces réseaux seront très hétérogènes. Malheureusement, il n'y a pas à ce jour une solution efficace contre les problèmes inhérents à cette hétérogénéité tels que l'intéropérabilité, la gestion de mobilité, la sécurité et la garantie de la qualité de service (QdS). De plus, aucune architecture existante ne permet une transparence et une convergence complète entre les différentes technologies. Il est donc nécessaire de concevoir d'autres architectures permettant d'atteindre ces objectifs.

D'autre part, les usagers seront de plus en plus mobiles et exigeants en termes de QdS pour leurs applications (voix, données, multimédia, etc.) Ils voudront bénéficier d'une mobilité sans coupure et d'une continuité de leur session ou service lorsqu'ils effectuent une relève. Cette relève pourrait être simplement due à un changement du canal de communication ou de l'adresse de routage (IP). Nous nous intéressons aux relèves verticales qui seront de plus en plus fréquentes dans un environnement hétérogène car les relèves horizontales ont déjà été largement étudiées dans la littérature. Différents protocoles ont été proposés pour permettre à un nœud mobile (MN) de maintenir sa connectivité lors de ses déplacements à travers des réseaux distincts. Ces protocoles, en particulier ceux proposés par l'IETF (dont l'un des plus connu est *Mobile IPv6*) ont plusieurs points faibles (latence, perte de paquets et signalisation élevées) et ne sont pas capables d'assurer une mobilité sans coupure

aux usagers, par exemple pour des applications temps-réel telle que la téléphonie IP (*voice over IP* - VoIP).

La prise en compte des problèmes ci-dessus évoqués est très importante pour la conception et le déploiement des réseaux sans fil ou mobile de prochaine ou quatrième génération (SFPG/4G). Nous nous focaliserons donc aux problèmes clés suivants : intégration, intéropérabilité et mobilité afin d'assurer une meilleure QdS aux usagers de même que de meilleures performances réseau pour les opérateurs. Afin de résoudre ces problèmes, cette thèse est basée sur quatre articles. Le premier article porte sur l'analyse des performances des protocoles de gestion de mobilité existants tandis que les trois derniers contiennent nos différentes solutions aux problèmes de mobilité et d'intégration dans les réseaux SFPG/4G.

Le premier article effectue une analyse des performances de plusieurs protocoles de gestion de mobilité au niveau de la couche IP. Une nouvelle approche analytique, plus approfondie, pour l'évaluation des performances des protocoles de gestion de mobilité y est proposée. Le cadre que nous proposons prend en compte plusieurs facteurs et leur interaction pour une modélisation plus rigoureuse. L'analyse effectuée montre qu'aucun de ses protocoles surclasse tous les autres par rapport aux différentes métriques de QdS considérées. Un compromis est donc nécessaire en fonction des objectifs de déploiement et de la QdS qu'un opérateur veut offrir à ses abonnés.

Dans le second article, nous proposons une nouvelle architecture hybride d'intégration appelée, *Integrated InterSystem Architecture* (IISA) et un nouveau protocole de gestion de mobilité appelé, *Handoff Protocol for Integrated Networks* (HPIN) pour des réseaux sans fil hétérogènes. L'architecture proposée introduit une nouvelle entité appelée, *Interworking Decision Engine* (IDE) qui joue le rôle d'une tierce partie pour permettre l'intégration et l'interopérabilité des différentes technologies d'accès. Chaque opérateur désirant offrir une mobilité globale à ses abonnés n'établit qu'une seule entente de services (*service level agreement*) avec

l'IDE au lieu de le faire avec tous les autres opérateurs. On voit bien la réduction de la complexité et les économies qui en découlent. Le protocole HPIN introduit une anticipation de la relève au niveau IP en utilisant les informations de la couche liaison, dans le but de réduire les interruptions de service. Ce protocole effectue aussi la découverte des réseaux d'accès, le transfert de contexte et la gestion locale de la mobilité. L'évaluation des performances montre que le protocole et l'architecture proposés donnent des meilleurs résultats que ceux disponibles dans la littérature.

Le troisième article par contre propose un nouveau mécanisme de décision de relève qui prend en compte plusieurs facteurs tels que la puissance du signal, la bande passante, le coût monétaire ou prix, le profil des usagers et le délai de relève. Un tel mécanisme de décision de relève est plus approprié à cause de l'hétérogénéité des réseaux SFPG/4G. Le protocole de gestion de relève proposé défini des nouveaux messages qui permettent d'avoir un protocole unifié assurant aussi bien la découverte des réseaux d'accès, la relève rapide et locale. La présentation de la solution est suivie par une batterie de tests afin d'en évaluer les performances. La fonction de décision de relève proposée permet d'assurer une meilleure répartition de charges dans les différents réseaux d'accès et donc de garantir un meilleur débit aux usagers. Le délai de relève, la perte de paquets et le trafic de signalisation sont aussi minimisés.

Finalement, le quatrième article est une amélioration du protocole présenté dans le deuxième article et est appelé *enhanced Handoff Protocol for Integrated Networks* (eHPIN). Cette amélioration consiste à effectuer les opérations critiques d'une relève par exemple, l'établissement de tunnels de communication et la mise à jour des caches d'association par anticipation avant ou pendant la relève au niveau de la couche liaison des données. L'étude comparative montre que eHPIN permet de réduire énormement le délai de relève et la perte des paquets. Une mobilité sans coupure et une continuité de services peuvent donc être garanties aux usagers.

ABSTRACT

The coexistence of diverse but complementary architectures and wireless access technologies has raised much interest in integration and interworking in order to benefit of their respective potentials and advantages. Amongst different layers of TCP/IP stack, the Internet Protocol (IP) technology seems the most appropriate to enable this integration. In fact, it is the lowest common layer of various access technologies and it allows the support of applications in a cost-effective and scalable way. The IP technology is expected to become the core or backbone network of the fourth or next generation wireless networks (4G/NGWN). Evolution through this integration is one of the paths to 4G/NGWN design, rather than investing efforts into developing new radio interfaces and technologies. Heterogeneity of 4G/NGWN brings new challenges such as, architecture design, mobility management, quality of service (QoS) provisioning and security. Despite several architectures proposed in the literature to solve those issues, none of them allows complete convergence and transparency between different technologies. Thus, there is a crucial needs to develop new and efficient architectures to reach those goals.

On the other hand, users or mobile nodes (MNs) will become more and more mobile and have increasing demands about quality of service for their applications (e.g., voice, data, multimedia). In other words, users should benefit of seamless roaming and services continuity when they roam across different access networks, known as handoff. The latter may be due to link layer switching or a change of the IP address. The focus of this thesis is on vertical handoff, which will be more frequent in 4G/NGWN due to the heterogeneity of access networks. Note that, the horizontal handoff was intensively studied in the literature. Several protocols have been proposed in the literature in order to allow MNs to maintain their connectivity with network during the handoff. These protocols, particularly those proposed by the IETF whose Mobile IPv6 is the most known, have several shortcomings (e.g.,

handoff latency, packet loss, and signaling overhead) and cannot provide seamless roaming to MNs, for example, for real-time applications such as IP Telephony or voice over IP (VoIP).

The issues aforementioned are crucial for the design and deployment of NGWN. Hence, we are interested in the following aspects : integration, interworking, and mobility management in order to guarantee appropriate QoS to users as well as better network performances for operators. In order to solve those issues, this thesis is based on four peer-reviewed journal papers. The first paper carries out performance analysis of several IP-based mobility management protocols. A novel analytical framework is proposed for performances evaluation of these protocols. The proposed framework takes into account the effect of several factors such as subnet residence time, packet arrival rate and wireless link delay for an effective modeling. The analysis performed shows that none of those protocols is better for all scenarios according to the QoS metrics considered. According to deployment goals and QoS level an operator wishes to provide to its subscribers, there is a certain level of trade-off required.

In the second paper, a novel integrated hybrid architecture called, *Integrated InterSystem Architecture* (IISA), and a new IP-based mobility management protocol called, *Handoff Protocol for Integrated Networks* (HPIN), are proposed for NGWN/4G. The proposed IISA architecture introduces a novel entity called, *Interworking Decision Engine* (IDE), which acts as a third-party between various access networks and technologies, and allows their integration and interworking. Each network operator or services provider which wants to allow global roaming to its subscribers establishes only one service level agreement (SLA) with the IDE manager rather than with all potential other operators. This allows significant cost and complexity reduction. The proposed HPIN scheme introduces IP-layer handoff anticipation by using link layer information in order to reduce services disruption. Moreover, HPIN performs also access networks discovery, local mobility manage-

ment and context transfer. Performance evaluation shows that IISA and HPIN give better results compared to works available in the open literature.

The third paper proposes a new handoff decision function which takes into account several parameters or factors such as monetary cost or price, bandwidth, signal strength, system performance, user preferences or profile, and handoff latency. Such handoff decision scheme is more appropriate in NGWN/4G due to the heterogeneity of access networks. On the other hand, the proposed handoff protocol defines new messages which enable design of a one-suite protocol that performs network selection, fast handoff, localized mobility management, context transfer and access networks discovery. The presentation of our solution is followed by performance evaluation which shows better results compared to others protocols. The handoff decision function allows a better load balancing amongst various access networks, thus, allows provisioning of higher data rate or throughput to users and lower packet loss rate.

Finally, the fourth paper is an improvement of the protocol presented in the second paper and is called, *enhanced Handoff Protocol for Integrated Networks* (eHPIN) that enables seamless services continuity and QoS guarantees for real-time applications in heterogeneous wireless environments. In eHPIN, the handshake procedure of all time consuming operations such as access router discovery, handoff anticipation, cache association update, and bi-directional tunnels setup are performed before or during the link layer switching (L2 handoff). Performance evaluation results shows that eHPIN allows better performance and significant reduction of handoff latency, signaling overhead cost, and packet loss compared to existing IPv6-based mobility management schemes. Thus, a seamless mobility and services continuity may be guaranteed to mobile users with eHPIN.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES SIGLES ET ABRÉVIATIONS

3G/4G :	Third/Fourth Generation
3GPP :	Third Generation Partnership Project
3GPP2 :	Third Generation Partnership Project 2
1xEV-DO/DV :	1x Evolution-Data Optimized/Data and Voice
AAA :	Authentication, Authorization and Accounting
AEN :	Access Edge Node
BAck :	Binding Acknowledgment
BEN :	Border Edge Node
BU :	Binding Update
CARD :	Candidate Access Router Discovery
CN :	Correspondent Node
CoA :	Care-of Address
CTD :	Context Transfer Data
CTDR :	Context Transfer Data Reply
CTReq :	Context Transfer Request
CXTP :	Context Transfer Protocol
DAD :	Duplicate Address Detection
eHPIN :	enhanced Handoff Protocol for Integrated Networks
FBAck :	Fast Binding Acknowledgment
FBU :	Fast Binding Update
F-HMIPv6 :	Fast Handoff for Hierarchical Mobile IPv6
FMIPv6 :	Fast Handovers for Mobile IPv6
FNA :	Fast Neighbor Advertisement
GGSN :	Gateway GPRS Support Node
GPRS :	General Packet Radio Service

HA :	Home Agent
HAck :	Handoff Acknowledgment
HI :	Handoff Initiate
HLR/HSS :	Home Location Register/Home Subscriber Server
HMIPv6 :	Hierarchical Mobile IPv6
HOReq/Rep :	Handoff Request/Reply
HPAck :	Handoff Preparation Acknowledgment
HPIN :	Handoff Protocol for Integrated Networks
HPN :	Handoff Preparation Notification
HPReq/Rep :	Handoff Preparation Request/Reply
HSPA :	High Speed Packet Access
IDE :	Interworking Decision Engine
IETF :	Internet Engineering Task Force
IEEE :	Institute of Electrical and Electronics Engineers
IISA :	Integrated InterSystem Architecture
IP (v4/v6) :	Internet Protocol (version 4/6)
LBAck :	Local Binding Acknowledgment
LBU :	Local Binding Update
LCoA :	on-Link Care-of Address
LTE/SAE :	Long Term Evolution/System Architecture Evolution
MAP :	Mobility Anchor Point
MIP(v4/v6) :	Mobile IP (version 4/6)
MN :	Mobile Node
NAR/PAR :	New/Previous Access Router
NCoA/PCoA :	New/Previous Care-of Address
NGWN :	Next Generation Wireless Networks

NLA :	New Link Attachment
NLAck :	New Link Attachment Acknowledgment
NLCoA/PLCoA :	New/Previous on-Link Care-of Address
PCF :	Packet Control Function
PDSN :	Packet Data Serving Node
PrRtAdv :	Proxy Router Advertisement
QdS (QoS) :	Qualité de Service (Quality of Service)
RA :	Router Advertisement
RCoA :	Regional Care-of Address
RIX Req/Rep :	Router Information eXchange Request/Reply
RS :	Router Solicitation
RtSolPr :	Router Solicitation for Proxy
SFPG :	Sans Fil de Prochaine Génération
SGSN :	Serving GPRS Support Node
SIP :	Session Initiation Protocol
SLA :	Service Level Agreement
SLRA :	Service Level and Roaming Agreement
TCP :	Transmission Control Protocol
UMB :	Ultra Mobile Broadband
UMTS :	Universal Mobile Telecommunication Systems
UTRAN :	UMTS Terrestrial Radio Access Network
VoIP :	Voice over IP
WCDMA :	Wideband Code Division Multiple Access
WIG :	WLAN Interworking Gateway
WiMAX :	Worldwide Interoperability for Microwave Access
WLAN :	Wireless Local Area Network

CHAPITRE 1

INTRODUCTION

La mobilité dans les réseaux de communication a permis de s'affranchir du mode de communication filaire traditionnel en apportant, non seulement des nouveaux services, mais aussi d'énormes défis de conception et de déploiement. Elle a une grande influence sur la performance des réseaux de communication. La tendance actuelle dans les réseaux de communication est la migration à des techniques de routage basées sur la commutation de paquets. La technologie IP (*Internet Protocol*) sera alors utilisée de bout-en-bout pour le routage ; on parle alors du concept du *tout-IP*. Les réseaux sans fil de prochaine ou quatrième génération (SFPG/4G) seront entièrement basés sur la technologie IP permettant ainsi des réels avantages par rapport aux différents systèmes de communication existants tels que, WLAN/IEEE802.11, UMTS/HSPA, CDMA2000 ou 1xEV-DO/DV et WiMAX ou ceux en cours de définition : LTE/SAE et UMB qui sont respectivement les évolutions des réseaux 3GPP et 3GPP2.

En effet, contrairement aux systèmes cellulaires de troisième génération (3G) dont l'objectif est de supporter différentes classes de services et des applications multimédia, le but des réseaux SFPG/4G est d'intégrer de façon transparente les systèmes sans fil courants (Hui & Yeung, 2003). La coexistence de ces différentes technologies d'accès sera le fondement du déploiement des réseaux SFPG. De cette hétérogèneité, il en résultera plusieurs problèmes tels que l'intégration, la gestion de mobilité et la sécurité. La gestion de mobilité au niveau IP a comme avantage l'indépendance vis-à-vis de la technologie radio utilisée et une certaine transparence, car elle est la plus basse couche de la pile TCP/IP commune aux différentes tech-

nologies des réseaux d'accès. Dans le cadre de cette thèse, nous nous focaliserons aux deux premiers problèmes c'est-à-dire, la gestion de mobilité et l'intégration.

1.1 Définitions et concepts de base

Les réseaux SFPG/4G permettront d'offrir et de transporter simultanément la voix, les données, la vidéo et le trafic multimédia. En d'autres termes, ils supporteront aussi bien les services et applications temps-réel et non temps-réel. Ils permettront une totale extension des réseaux sans fil actuels en termes d'architectures, de services, d'applications, de capacité et d'hétérogénéité. Avec l'introduction des services multimédia dans un environnement sans fil et mobile, les usagers solliciteront l'accès à des technologies sans fil à large bande passante et auront des exigences élevées de qualité de service (QdS), semblables à celles des réseaux fixes. Les réseaux SFPG/4G seront basés sur des technologies d'accès diverses ce qui permettra de tirer profit de leur complémentarité au lieu de définir une nouvelle technologie radio (Hui & Yeung, 2003). Une illustration de l'hétérogèneité d'un système de communication SFPG/4G est donnée à la Figure 1.1.

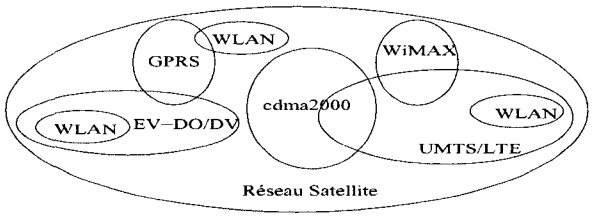

Figure 1.1 Architecture typique des réseaux SFPG/4G.

Au moment du déploiement des réseaux SFPG/4G, il est prévu que la version 6 du protocole IP (IPv6) définie dans Deering & Hinden (1998) soit largement utilisée. De plus, les deux groupes de standardisation des réseaux sans fil de troisième génération, 3GPP et 3GPP2, ont adopté cette technologie pour le routage (Chen

& Zhang, 2004). Ainsi, dans cette thèse, IPv6 est considéré comme technologie de base pour l'intégration et la transparence entre les différents réseaux d'accès. Parmi les principales caractéristiques des réseaux SFPG/4G, on peut citer (Hui & Yeung, 2003; Tafazolli *et al.* , 2005) :

- utilisation des terminaux multi-modes avec plusieurs interfaces, qui sont capables de supporter différentes technologies d'accès (*multihoming*) ;
- connectivité en tout temps et en tout lieu (*any time and anywhere*), faisant ainsi référence aux concepts de nomadisme et d'ubiquité ;
- réseaux d'accès intégrés avec une dorsale commune basée sur le protocole IP ;
- un environnement sécurisé dans lequel les entités réseaux sont déployées et interagissent ;
- support des services de voix, de données, multimédia et personnalisés à un coût minimal. Ces services peuvent être accessibles à partir des réseaux appartenant à des fournisseurs ou opérateurs différents.

Le terme mobilité dans les réseaux de communication réfère à la possibilité d'accéder à des services indépendamment de la localisation et du déplacement de l'usager. Trois types de mobilité sont à considérer dans les réseaux SFPG/4G : la mobilité terminale, la mobilité personnelle et la mobilité ou portabilité des services (Tafazolli *et al.* , 2005). La mobilité terminale réfère à la capacité de localiser et d'identifier un terminal ou nœud mobile, en permettant à ce dernier d'accéder aux services offerts par le réseau à partir de n'importe quelle position lors de ses déplacements à travers différents accès radio. Un numéro d'identification, par exemple une adresse IP, est associé à chaque terminal.

Par contre, la mobilité personnelle implique l'identification des usagers auxquels elle permet, non seulement de recevoir et d'initier des appels, mais aussi d'accéder aux services de réseau à partir de n'importe quel terminal et de n'importe quelle position de façon transparente. L'usager dispose d'un numéro d'identification per-

sonnel ayant généralement pour support une carte à puce. Enfin, la mobilité des services réfère à la capacité du réseau d'identifier les usagers en mouvement, de permettre à ces usagers d'initier et de recevoir des appels, et de fournir les services auxquels les usagers ont souscrit en fonction de leur localisation. Dans le cadre de notre recherche, nous nous intéressons plus particulièrement à la mobilité terminale.

La gestion de mobilité dans les systèmes de communication sans fil en particulier la mobilité terminale est basée sur deux aspects clés à savoir la gestion de relève (*handoff management*) et la gestion de localisation (*location management*). Ces deux procédures engendrent un trafic de signalisation important dans le réseau, d'où l'intérêt de trouver des mécanismes robustes et efficaces afin de minimiser ce trafic. La gestion de localisation permet au système de connaître à tout moment la position courante du nœud mobile (MN). La relève (*handover* ou *handoff*) représente le transfert automatique du canal de communication d'un point d'accès à un autre lors d'un changement de réseau du MN ou quand la qualité du signal se dégrade.

Dans les réseaux SFPG, la relève peut aussi être due aux préférences et au profil de l'usager. Ainsi, une relève peut être d'une part obligatoire ou forcée et d'autre part volontaire (Nasser *et al.* , 2006). La gestion de relève doit assurer la continuité de la session en cours ou tout autre type de lien entre le MN et le réseau. La procedure de relève se divise en trois phases : la détection, la décision et l'exécution. Au niveau IP, on distingue deux principaux types de relèves : la *relève intra-domaine* (associée à la *micro-mobilité*) et la *relève inter-domaine* (associée à la *macro-mobilité*). Lorsque le MN en cours de relève demeure dans le même domaine, on parle de relève intra-domaine. Par contre, la relève inter-domaine a lieu lorsque les deux points d'accès appartiennent à des domaines différents.

Il existe trois approches pour une décision de relève dépendemment de la contribution du réseau ou du nœud mobile. En effet, on peut avoir une relève contrôlée

par le réseau (*network-controlled handoff*), une relève contrôlée par le nœud ou terminal mobile (*mobile-controlled handoff*) et une relève assistée (*mobile-assisted handoff*). La relève contrôlée par le réseau est une solution centralisée dans laquelle le réseau décide d'une relève à partir des mesures effectuées sur la puissance du signal reçu par les nœuds mobiles. Les inconvénients majeurs de cette approche sont d'une part les exigences en terme de puissance de calcul et le maintien de l'information en une seule entité centrale et d'autre part l'absence d'une connaissance exacte des conditions courantes au niveau de chaque nœud mobile. Dans la version contrôlée par le nœud mobile, ce dernier a l'intelligence et l'autorité de choisir le point d'attache au réseau à partir de ses propres mesures.

Cette approche peut avoir un impact sur plusieurs facteurs tels que la stabilité du réseau, la sécurité et l'équité, car elle est distribuée et aucune politique globale ne peut être appliquée. Avec la relève assistée, le nœud mobile effectue plusieurs mesures sur certains facteurs, par contre la décision de relève est faite au niveau réseau. Dans ce cas, les conditions courantes au niveau du nœud mobile peuvent être prises en compte, mais le réseau fait toujours face à une surcharge du trafic de signalisation et aux exigences de traitement des informations. Nonobstant l'existence de ces trois approches, sur le plan pratique, dans les réseaux sans fil et mobile courants, l'usager n'a pas le contrôle sur la procédure de relève ; on parle alors de *relève passive*. Par ailleurs, dans les réseaux SFPG/4G, l'usager ou mieux le terminal aura la possibilité de décider à quel moment une relève peut être effectuée. On parle dans ce cas de *relève proactive* (Nasser *et al.* , 2006).

D'autre part, avec la coexistence de plusieurs technologies d'accès dans les réseaux SFPG, la gestion de la mobilité radio demeure préoccupante. En effet, un nœud mobile pourrait changer de technologie d'accès durant ses déplacements ou en fonction de ses préférences. Ainsi, on peut distinguer deux types particuliers de relèves à savoir, la *relève horizontale* (ou *intra-système* ou encore *intra-technologie*)

et la *relève verticale* (ou *inter-système* ou encore *inter-technologie*). La première a lieu quand la technologie utilisée dans les deux domaines ou réseaux est la même tandis que la seconde se produit dans le cas contraire. Une relève verticale se produit fréquemment dans un environnement sans fil avec chevauchement (partiel ou complet) des zones de couverture de différents réseaux (systèmes).

En outre, elle est souvent asymétrique et peut être subdivisée en relève verticale montante et descendante (Stemm & Katz, 1998). La première se produit lors d'une relève d'un réseau ayant une petite couverture mais offrant une bande passante élevée vers un réseau de large couverture mais ayant une bande passante plus petite. C'est le cas d'une relève d'un réseau WLAN vers un réseau cellulaire 3G. Par contre, une relève verticale descendante se produit dans l'autre sens. Notons que la définition de cette asymétrie doit être revue lorsqu'on considère la technologie WiMAX comme réseau d'accès. La conception d'un mécanisme pour gérer une relève verticale devrait donc avoir comme objectif :

- minimiser la latence de relève, le gaspillage de la puissance ou énergie de la batterie, l'interférence, le nombre de relèves et le trafic additionnel utilisé pour supporter ces relèves ;
- maximiser la fiabilité et la performance, autrement dit, une session en cours doit maintenir une bonne qualité après l'exécution de la relève ;
- maintenir une mobilité sans coupure (c'est-à-dire, minimiser l'interruption de service causée par les relèves) et une répartition adéquate des charges afin de réduire la probabilité de blocage des nouvelles sessions ou celles en cours de relève.

La Figure 1.2 illustre une relève horizontale entre les systèmes WLAN1 et WLAN2, et une relève verticale entre les systèmes WLAN3 et UMTS.

La notion de qualité de service (QdS) semble parfois difficile à définir. Cependant,

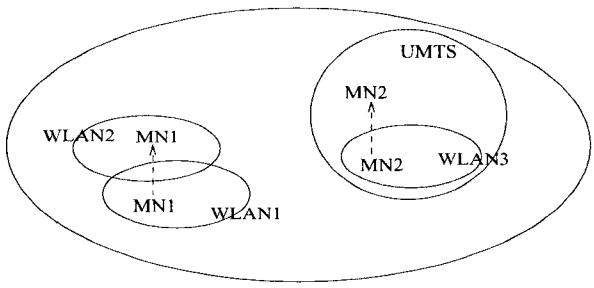

Figure 1.2 Relèves horizontale et verticale.

on peut dire que la QdS pour un usager représente son degré de satisfaction pour un service auquel il a souscrit. Elle consiste à fournir un service conforme aux exigences des usagers. Dans un réseau de communication mobile et sans fil, la garantie de la QdS demeure un défi majeur, en particulier pour des applications multimédia qui ont des contraintes strictes de délai. La mobilité des usagers peut provoquer une interruption de service à cause de plusieurs facteurs. Par exemple, lors d'une relève, le délai de rétablissement de la connexion peut être assez élevé. Pour ce faire, il faut garantir une mobilité sans coupure (*seamless roaming*) et une continuité de service aux usagers. Dans les réseaux sans fil et mobile, la QdS peut être définie à partir de plusieurs métriques telles que le délai ou latence de relève, le taux de perte de paquets, le trafic de signalisation, la probabilité de blocage, le débit et le délai de bout-en-bout. Ainsi, garantir une mobilité sans coupure reviendrait par exemple à minimiser le délai de relève, le taux de perte de paquets, le trafic de signalisation et la probabilité de blocage.

L'intervalle de temps durant lequel un nœud mobile ne peut pas recevoir ou transmettre de paquets durant une relève est appelé latence de relève (*handoff latency*). Ce temps a une influence sur la performance de la relève en particulier pour des applications temps-réel. Dans un environnement IP mobile, deux types de signalisation sont liés aux mises à jour des associations (*binding update*) : celles faisant suite à l'acquisition d'une nouvelle adresse et celles nécessaires au rafraîchis-

sement des caches d'associations (*binding refresh*). Ces deux procédures de mise à jour génèrent un trafic de signalisation. Les usagers mobile sont plus sensibles à l'interruption de leur session ou communication en cours qu'à son rejet lors de l'initialisation. La probabilité de blocage exprime donc le taux qu'une session soit interrompue prématurément suite à un échec de relève. Le blocage d'une session peut être causé par plusieurs facteurs parmi lesquels, la latence de relève, la détérioration du rapport signal à bruit ou l'indisponibilité du canal de communication. Le taux d'échec de relève doit être maintenu en dessous d'un certain seuil. Afin d'assurer une connectivité entre différentes entités dans les réseaux SFPG/4G, il est nécessaire d'avoir une convergence ou interopérabilité de leurs fonctionalités.

1.2 Éléments de problématique

Bien que les réseaux SFPG/4G soient encore à la phase de conception et que leur déploiement soit envisagé vers 2010 (Roberts *et al.* , 2006), plusieurs travaux sont en cours pour faire face aux défis qu'ils présentent. Ces réseaux consisteront en une intégration des différentes technologies existantes qui peuvent être complémentaires entre elles. Malgré, leur potentiel, aucun des réseaux sans fil et mobile existants n'est capable d'offrir simultanément une bande passante élevée, une couverture à grande échelle, des délais de livraison des paquets faibles, etc. La complémentarité de ces réseaux suscite l'intérêt de leur intégration et interopérabilité. De cette intégration, il en résultera un système sans fil hétérogène apportant de nouveaux défis dans la conception de l'architecture et des protocoles, le support de la mobilité et la garantie de la qualité de service.

Une ébauche des défis des réseaux SFPG/4G est faite par Hui & Yeung (2003), de même que quelques pistes de solutions proposées dans la littérature y sont présentées. Par ailleurs, malgré les travaux effectués, plusieurs questions demeurent

ouvertes. Avec la mobilité, les usagers feront face aux relèves tant intra-technologie (horizontale) qu'inter-technologie (verticale) dans un environnement hétérogène multi-accès. Il est essentiel que les applications s'exécutant sur les terminaux ou nœuds mobile ne soient pas perturbées. En fait, qu'elles n'aient pas à se soucier du mouvement des usagers. Le système doit assurer une communication sans interruption avec une dégradation minimale de la qualité de service pour les usagers mobile. Cela peut être obtenu par une réduction de la latence et de la probabilité d'échec de relève en dessous de certains seuils. Par exemple, pour la voix sur IP (VoIP), le taux de perte de paquets acceptable est de 3% (Vivaldi *et al.* , 2003).

Plusieurs protocoles de gestion de mobilité ont été proposés dans la littérature et sont souvent spécifiques à chaque couche de la pile TCP/IP et opèrent indépendamment. Au niveau IP, afin de gérer la mobilité, l'*Internet Engineering Task Force* (IETF) a proposé le protocole *Mobile IPv6* (MIPv6) (Johnson *et al.* , 2004) et plusieurs de ses extensions afin de permettre aux nœuds mobile de maintenir leur connexion active lors de leurs déplacements. Toutefois, plusieurs problèmes subsistent avec ces protocoles. La mise à jour des informations du nœud mobile chaque fois qu'il change de position ou obtient une nouvelle adresse temporaire induit un gaspillage des ressources. Une approche d'anticipation de la relève a été introduite pour améliorer la performance de MIPv6 en utilisant par exemple les informations de la couche liaison de données.

Cependant, cette anticipation nécessite une bonne précision pour détecter et prédire les déplacements du nœud mobile sans qu'il ne soit trop complexe. Bien que cette approche permette de réduire la latence de relève, des problèmes de synchronisation apparaîssent, de même qu'une augmentation du trafic de signalisation. Une autre approche pour améliorer les performances consiste en une subdivision de la mobilité (intra-domaine et inter-domaine) en introduisant une hiérarchisation. Toutefois, aucune de ces extensions ne permet d'avoir simultanément un trafic de si-

gnalisation faible, une perte minimale des paquets et une latence de relève moindre (Pérez-Costa *et al.* , 2003; Gwon *et al.* , 2004). Ainsi, aucun d'entre eux ne permet d'assurer une mobilité sans coupure dans le contexte des réseaux SFPG/4G.

Dans un environnement multi-accès, le choix du réseau qui desservira l'usager est un défi considérable, car chacune des technologies possède ses propres caractéristiques. Traditionnellement, ce choix est basé sur la puissance du signal reçu et la disponibilité du canal de communication. Par contre, ces critères ne peuvent s'appliquer efficacement que dans le cas d'une relève horizontale. Dans les réseaux SFPG, les relèves verticales seront fréquentes, alors il y a lieu de définir de nouveaux mécanismes de sélection de réseaux (interfaces) qui devront prendre en compte plusieurs critères tels que la puissance du signal, les conditions du réseau, les préférences des usagers et le coût monétaire ou prix. La sélection du réseau ne devrait pas être limitée en utilisant uniquement les informations disponibles au niveau du réseau d'accès telles que la puissance du signal, la charge ou la disponibilité de la bande passante. Il faudra prendre en compte les informations et exigences aussi bien au sein du réseau d'accès, du réseau cœur ou nominal et du terminal mobile.

Il est difficile de décider du moment opportun où se produit une relève dans un environnement hétérogène en raison des particularités des mécanismes de relève dans chacun des systèmes. D'autre part, pour un nœud mobile équipé de plusieurs interfaces, la manière évidente pour découvrir un réseau d'accès est de maintenir toutes les interfaces continuellement actives. Cependant, une telle approche engendrera aussi bien une consommation excessive de l'énergie du terminal mobile que les ressources du réseau, même si aucun paquet n'est envoyé ou reçu par le terminal. Il est donc nécessaire de trouver une approche plus efficace pour activer les interfaces afin de garantir un meilleur compromis entre la découverte de réseau et la consommation d'énergie. Des investigations pour des nouvelles stratégies de relève dans un environnement hétérogène sont donc requises.

Le problème d'intégration et d'interopérabilité des réseaux de communication a été largement étudié. Deux modèles d'architecture, *loose* et *tight coupling*, ont été proposés pour l'intégration des réseaux cellulaires 3G et WLAN (3GPP, 2004; 3GPP2, 2004). De plus, six scénarios sont définis dans 3GPP (2004) et 3GPP2 (2006) afin de supporter l'interopérabilité. Cependant, tous ces scénarios ne sont pas encore totalement garantis et les deux architectures proposées présentent un certain nombre de faiblesses. En effet, avec le couplage *tight*, un réseau WLAN apparaît comme un réseau d'accès pour le réseau cœur 3G. Ainsi, le trafic du réseau WLAN sera acheminé vers le réseau 3G, ce qui pourrait causer des problèmes de capacité, car ce dernier n'est pas adapté au trafic à haut débit. Par ailleurs, pour le couplage *loose*, les différents réseaux sont déployés indépendamment, permettant de réduire la complexité et les coûts. Cependant, il est difficile de garantir la continuité de services lors d'une relève due à la latence et la perte des paquets qui sont assez élevées. Ces deux modèles d'intégration sont strictement limités aux réseaux WLAN et cellulaires 3G. D'autre part, une architecture d'intégration devrait respecter certaines exigences telles que, être économique, évolutive, assurer une mobilité sans coupure et garantir un niveau adéquat de sécurité (Akyildiz *et al.* , 2005).

Une des solutions courantes pour permettre la mobilité des usagers est l'établissement des accords de services et d'itinérance (*Service Level and Roaming Agreements* - SLA/RA) entre les opérateurs et fournisseurs de services. Toutefois, cette approche présente des limites. En effet, les opérateurs de téléphonie ne sont pas disposés à donner accès à leur base de données aux autres opérateurs même en présence d'un SLA. Or l'accès aux bases de données est nécessaire pour certaines opérations par exemple, l'authentification et la facturation ou AAA (*Authentication, Authorization and Accounting*). D'autre part, le nombre d'opérateurs étant très élevé, de surcroît avec la libéralisation du marché des télécommunications, il devient presque impossible à un opérateur d'avoir des accords SLA/RA directement

avec tous les autres opérateurs. En effet, ça serait très coûteux et exigeant.

De plus, si un usager effectue une relève entre deux réseaux n'ayant pas de SLA/RA, il ne peut pas bénéficier du maintien de la connexion de sa session en cours, même s'il est abonné indépendamment aux deux réseaux. Il en découle la nécessité d'une architecture permettant l'itinérance globale des usagers, qui ne soit pas basée sur des accords SLA/RA directs entre opérateurs. Au sein de certains groupes, tel que le *GSM Association*[1], des dispositions ont été prises pour permettre l'itinérance entre opérateurs dont les réseaux sont basés sur la technologie GSM. Mais dans la pratique, il y a encore des conflits opérationnels et le problème d'itinérance demeure ouvert, en particulier entre deux technologies distinctes.

1.3 Objectifs de recherche

L'objectif principal de cette thèse est de proposer des mécanismes efficaces de gestion de mobilité dans les réseaux sans fil de prochaine génération (SFPG/4G) offrant des garanties de qualité de service (QdS) aux usagers ainsi que l'intégration et l'interopérabilité des réseaux de communication tant filaire que sans fil et mobile existants. Plus spécifiquement, cette thèse vise les objectifs suivants :

- analyser les protocoles et mécanismes proposés dans la littérature pour la gestion de mobilité avec support de la QdS dans les réseaux sans fil et mobile basés sur la technologie IP afin d'en déceler les faiblesses et les problèmes qui ne sont pas encore résolus adéquatement ;
- proposer une architecture permettant l'intégration et l'interopérabilité des différents réseaux existants dans un environnement hétérogène multi-accès ;
- concevoir des nouveaux mécanismes pour la sélection de réseau, la gestion de

[1]GSM Association : htttp://www.gsmworld.com

relève, le support de la QdS et la localisation permettant une itinérance sans coupure des sessions ou services des usagers durant leurs déplacements ;

- évaluer l'efficacité de cette architecture et la performance des mécanismes proposés en tenant compte des exigences et spécifications des réseaux SFPG/4G. Cette évaluation sera basée sur une comparaison avec les travaux existants qui abordent les mêmes problèmes.

1.4 Esquisse méthodologique

Une démarche méthodologique basée sur une approche analytique et par simulation servira de guide pour la validation des différentes contributions de cette thèse. Nous commencerons par une analyse approfondie des caractéristiques des réseaux SFPG/4G. Cette phase sera effectuée à partir d'une revue de littérature pertinente qui nous permettra d'identifier les enjeux de la mobilité et les exigences sur la QdS dans les réseaux SFPG/4G. Ceci nous permettra de déceler les avantages et faiblesses des approches et solutions disponibles dans la littérature. Ensuite, une architecture intégrant différents réseaux dans un environnement mobile et sans fil hétérogène sera proposée.

Pour atteindre cet objectif, nous analyserons les architectures d'intégration disponibles dans la littérature afin d'en dégager les principales caractéristiques et requis pouvant nous permettre de concevoir une nouvelle architecture hybride. Cette dernière devra permettre une transparence de l'hétérogènéité et une interopérabilité harmonieuse. De plus, nous nous baserons sur les recommandations et spécifications des organismes tels que l'IETF, le 3GPP/3GPP2, l'IEEE et l'ETSI pour la conception de cette architecture qui devra utiliser autant que possible les entités existantes et minimiser l'ajout des nouvelles. L'évolutivité et la fiabilité doivent être prises en compte comme critère de performance d'une telle architecture pour faire face aux

problèmes de survabilité, de fiabilité et de tolérance aux pannes.

Une stratégie intelligente et efficace pour la sélection du meilleur réseau (interface) disponible auquel un nœud mobile se connecte sera définie. Cette approche de sélection permettra d'assurer une meilleure répartition de charge dans les réseaux d'accès et un meilleur contrôle d'admission. Nous allons concevoir de nouveaux mécanismes de gestion de relève qui permettront de garantir une mobilité sans coupure et une continuité des services, par exemple en minimisant la perte de paquets, le trafic de signalisation, le délai de relève et la probabilité de blocage ou d'échec de relève. Une approche modulaire sera utilisée pour développer les différents mécanismes ci-dessus mentionnés, afin de permettre une flexibilité d'intégration dans la solution globale. L'impact de chaque mécanisme proposé sera étudié.

L'évaluation des performances de l'architecture et des mécanismes proposés sera effectuée. Comme outils d'évaluation, les logiciels OPNET et MATLAB seront utilisés. Autrement dit, les différents mécanismes proposés seront implémentés dans ces deux logiciels. Plusieurs tests seront effectués avec différents scénarios (par exemple type de trafic, modèle de mobilité) afin de valider les solutions proposées. Une étude comparative sera aussi faite avec les autres propositions disponibles dans la littérature. Les métriques et critères définis précédemment seront utilisés pour l'évaluation des performances.

1.5 Principales contributions et originalité

Les principales contributions de cette thèse qui permettent de faire avancer la recherche sur les deux défis majeurs des réseaux SFPG/4G à savoir, la gestion de mobilité et l'intégration des réseaux, sont au nombre de quatre : un modèle analytique pour évaluer les performances des protocoles de mobilité, une architecture

hybride intégrant les réseaux de communication, un mécanisme de décision de relève et des protocoles de gestion de mobilité. Elles peuvent être résumées comme suit :

- *Modèle analytique d'évaluation de performances* : la proposition d'un nouveau protocole doit être accompagnée d'une preuve de concepts. Pour y arriver, quatre approches peuvent être utilisées : simulation, modélisation analytique, validation formelle et prototypage (*testbed*). Les deux premières sont les plus courantes dans la littérature. En ce qui concerne la modélisation analytique des performances des protocoles de gestion de mobilité, les approches abordées dans la littérature sont très simplifiées et peu réalistes. En effet, l'interaction entre les différentes métriques (par exemple, le délai, le trafic de signalisation et la perte de paquets) est souvent ignorée. De même, l'influence de plusieurs facteurs tels que la mobilité des usagers, les conditions réseau et le type de trafic n'est pas prise en compte conjointement. Pour ce faire, nous proposons un modèle analytique plus robuste qui permet de prendre en compte l'interaction entre les métriques ainsi que l'influence des facteurs sur ces métriques. Une telle approche globale n'a pas été proposée auparavant, ce qui fait de notre travail une contribution originale.

- *Architecture hybride d'intégration* : afin de permettre l'intégration des réseaux de communication actuels, nous proposons une nouvelle architecture introduisant une tierce-partie qui permet d'assurer l'interopérabilité dans un environnement hétérogène. L'ajout de nouvelles entités réseau a été minimisé afin de garantir un déploiement économique. L'accent a été mis sur une extension des fonctionnalités des entités existantes. L'architecture proposée permet une séparation entre le trafic de signalisation et le trafic de données, ce qui allège la charge du tierce-partie. En outre, elle offre un meilleur compromis entre les deux modèles génériques disponibles dans la littérature. L'évaluation

de performances de cette architecture montre qu'elle vérifie tous les requis tels que la minimisation du coût des infrastructures et l'évolutivité.

- *Mécanisme de décision de relève* : la décision de relève basée sur la qualité de la puissance du signal ou la disponibilité de canal ou encore de bande passante n'est pas appropriée dans les réseaux SFPG. Pour faire face à ce problème, nous avons proposé une nouvelle approche basée sur une fonction de score qui prend en compte différents facteurs tels que la puissance du signal, la bande passante, le coût ou prix d'une session, la vitesse des usagers et le profil des usagers. En prenant en compte ces différents facteurs, l'usager bénéficie d'une meilleure connectivité. En effet, l'usager ou le nœud mobile sera connecté au réseau qui permet de maximiser sa fonction de score. Ce mécanisme incorpore aussi une stratégie de gestion des interfaces afin de garantir un meilleur compromis entre la découverte de réseaux d'accès et la consommation de l'énergie des terminaux mobile. Cette contribution est non seulement originale mais très importante pour les réseaux SFPG dû à l'hétérogénéité des réseaux d'accès. Une telle fonction de décision de relève permet d'avoir une idée sur l'impact de chaque facteur de même qu'une meilleure répartition de charges à travers les réseaux d'accès disponibles.

- *Protocoles de gestion de mobilité* : plusieurs protocoles de gestion de mobilité sont disponibles dans la littérature avec leurs avantages et incovénients. Ces protocoles sont développés séparemment et tentent de résoudre un problème spécifique de mobilité. Nous proposons différentes versions de protocole de gestion de mobilité ; ce qui constituent une amélioration considérable des protocoles proposés par l'IETF. Les protocoles proposés ne traitent pas seulement la mobilité IP, mais utilisent l'information des autres couches pour offrir des meilleures performances et assurer une meilleure qualité de service aux

usagers. Le mécanisme de décision de relève ci-dessus mentionné est incorporé de même que des mécanismes plus intelligents pour la découverte des réseaux d'accès, l'anticipation de la relève et le transfert de contexte. Un nouveau mécanisme de mise à jour des caches d'association est proposé afin de réduire le trafic de signalisation sur la portion sans fil du réseau. Ces protocoles permettent d'assurer une mobilité sans coupure et la continuité de services dans les réseaux SFPG/4G. Ces deux exigences étant cruciales dans ces réseaux.

1.6 Plan de la thèse

Le reste de cette thèse est organisée comme suit. Le Chapitre 2 présente une revue critique et sélective de la littérature sur deux problèmes clés des réseaux SFPG/4G à savoir, l'intégration des différents systèmes de communication et la gestion de mobilité. Les deux principales architectures génériques d'intégration disponibles dans la littérature sont abordées ainsi que les défis et problèmes qui en découlent. Par la suite, les protocoles de gestion de mobilité dans les environnements sans fil et mobile hétérogènes sont présentés. Une analyse approfondie de ces protocoles est aussi effectuée, afin d'identifier les problèmes qui subsistent encore. Ayant opté pour une thèse par articles, les Chapitres 3 à 7 contiennent respectivement les différents articles qui décrivent notre contribution.

Plus précisément, le Chapitre 3 présente l'article intitulé *An Analytical Framework for Performance Evaluation of IPv6-based Mobility Management Protocols* qui a été accepté pour publication dans la revue *IEEE Transactions on Wireless Communications*. Dans cet article, nous proposons une nouvelle approche analytique, plus approfondie, pour évaluer les performances des protocoles de gestion de mobilité. Le cadre que nous proposons prend en compte plusieurs facteurs et leur interaction pour une modélisation plus rigoureuse. L'article intitulé *An Architec-*

ture for Seamless Mobility Support in IP-based Next-Generation Wireless Networks accepté pour publication dans la revue *IEEE Transactions on Vehicular Technology* est présenté au Chapitre 4. Dans cet article, nous proposons une nouvelle architecture hybride d'intégration et un nouveau protocole de gestion de mobilité dans un environnement sans fil hétérogène, qui étend les protocoles proposés par l'IETF.

Le Chapitre 5 intitulé *Adaptive Handoff Scheme for Heterogeneous IP Wireless Networks* est un article soumis à la revue *Computer Communications (Elsevier).* Un nouveau mécanisme de décision de relève, plus approprié aux réseaux SFPG/4G, basé sur une fonction de score y est proposé. Cette fonction prend en compte plusieurs facteurs tels que le coût ou prix d'une session, la bande passante disponible et la puissance du signal. De plus, un autre protocole de gestion de relève est proposé. Une amélioration du protocole décrit dans le Chapitre 4 est traîtée dans le Chapitre 6 qui présente l'article intitulé *Enhanced Fast Handoff Scheme for Heterogeneous Wireless Networks* soumis à la revue *Computer Communications (Elsevier).* Cette amélioration consiste à effectuer les opérations critiques d'une relève par exemple, l'établissement de tunnels de communication et la mise à jour des caches d'association par anticipation. Dans le Chapitre 7, une discussion générale sur les différents résultats obtenus et une synthèse de notre contribution sont faites. Enfin, cette thèse se termine par une conclusion qui permet d'effectuer un bilan en regard de nos objectifs de recherche, d'exposer les limites de notre contribution et évoquer les recommandations pour des travaux futurs.

CHAPITRE 2

INTÉGRATION, INTEROPÉRABILITÉ ET MOBILITÉ

La coexistence des réseaux de communication et technologies distincts constitue le fondement de la conception et du déploiement des réseaux sans fil de prochaine génération (SFPG/4G). Pour ce faire, l'intégration et l'interopérabilité de ces réseaux sont nécessaires pour tirer profit de leurs avantages respectifs. Toutefois, cette intégration apporte plusieurs défis auxquels il faut faire face. Parmi ces défis, on peut citer la gestion de mobilité et des ressources, la conception d'architectures et de protocoles, la sécurité et la garantie d'une meilleure qualité de service (QdS). Plusieurs travaux ont été entrepris dans la littérature afin de solutionner ces défis. Nonobstant les efforts consentis et les résultats obtenus, plusieurs problèmes subsistent. En effet, les activités de recherche sur les réseaux SFPG/4G restent d'actualité et plusieurs projets sont en cours. Dans ce chapitre, nous allons faire un survol critique et sélectif des travaux disponibles dans la littérature sur les défis énumérés ci-avant en particulier sur la conception d'architectures d'intégration, la gestion de mobilité et par conséquent la garantie de QdS dans les réseaux SFPG.

2.1 Mécanismes d'intégration

Les réseaux mobiles et sans fil courants peuvent être vus comme complémentaires les uns des autres bien qu'ils aient été conçus pour des besoins spécifiques. Les réseaux cellulaires 3G, tels que UMTS et 1xEV-DO, ont l'avantage d'offrir une large couverture (périmètre géographique plus étendu) tandis que leurs inconvénients se caractérisent par la capacité limitée de la bande passante et les coûts opérationnels

élevés. Par contre, la technologie WLAN, telle que IEEE 802.11, offre une large bande passante avec des coûts d'opérations faibles, bien que sa couverture soit moins étendue et ne peut supporter que des usagers ayant un taux de mobilité faible.

Le déploiement exponentiel des réseaux WLAN (Al-Gizawi *et al.* , 2002) justifie le fait qu'ils joueront un rôle clé dans la transmission des données dans le futur. Ce fait est bien connu des opérateurs de téléphonie mobile et ils essaient d'en tirer profit. Six scénarios ont été définis par les deux organismes de standardisation des réseaux cellulaires 3G (3GPP, 2004; 3GPP2, 2006) afin de supporter l'interopérabilité de leurs technologies avec les réseaux WLAN/IEEE 802.11. Au sein de l'IEEE plusieurs projets[1] sont en cours pour l'extension du standard IEEE 802.11 pour des besoins d'interopérabilité avec les autres technologies. Nonobstant, la richesse des contributions, plusieurs questions restent ouvertes et nécessitent donc que l'on s'y intéresse. Le groupe de travail IEEE 802.11u s'intéresse au développement d'un standard qui permettra l'interopérabilité d'un réseau IEEE 802.11 avec des réseaux externes auquel il est connecté. D'autre part, le groupe de travail IEEE 802.11r traite la spécification des transitions rapides entre différents BSS (*Basic Service Set*) dans le but d'offrir une relève sans coupure (*seamless handoff*).

Le standard IEEE 802.11r est une extension de IEEE 802.11f ou *Inter-Access Point Protocol* (IAPP) et le type d'application cible étant la voix sur IP (VoIP). Enfin, le standard IEEE 802.21 ou *Media Independent Handover* (MIH) fourni une intelligence à la couche liaison de données et d'autres informations appropriées du réseau aux couches supérieures afin d'optimiser la relève entre différentes technologies, c'est-à-dire basées sur IEEE 802.11 et n'importe quelle autre technologie par exemple, 3GPP et 3GPP2. Notons qu'au niveau radio, les deux techniques qui

[1]WLAN Working Group : htttp://www.ieee802.org/11/

ont été ajoutées, dans les réseaux 3GPP LTE/SAE et 3GPP2 UMB, aux techniques déjà deployées, CDMA, TDM et OFDM dans les réseaux cellulaires 3G sont OFDMA (*Orthogonal Frequency Division Multiple Access*), MIMO (*Multiple Input Multiple Output*) et SDMA (*Space Division Multiple Access*). L'objectif visé étant d'augmenter la performance, offrir une meilleure capacité, une plus grande couverture et une meilleure qualité pour les réseaux SFPG/4G. Ces techniques sont combinées afin d'avoir une seule interface radio en tirant profit des meilleurs aspects de chacune d'elles.

Étant donné qu'aucun des systèmes ne permet de satisfaire tous les scénarios en terme par exemple, de mobilité et de qualité de service, et avec les exigences des nouveaux services multimédia au niveau des usagers, il y a lieu de trouver de nouvelles approches pour la conception des réseaux SFPG/4G. Deux approches sont possibles pour cette conception (Akyildiz *et al.* , 2005). En effet, la première approche consisterait au développement d'un nouveau système sans fil en termes d'interfaces radio et de technologies qui permettrait de satisfaire les exigences de QdS des usagers. Ceci permettrait par exemple d'atteindre le débit envisagé supérieur à 100 Mbps pour les réseaux SFPG/4G. La seconde approche serait une intégration intelligente des systèmes sans fil existants afin de permettre aux usagers d'être servis par le meilleur système disponible. La première approche semble moins pratique car elle nécessite plus de temps de développement et de déploiement et serait par conséquent très coûteuse. La seconde approche semble plus réaliste (Hui & Yeung, 2003) et c'est cette dernière que nous privilégions dans cette thèse.

De cette intégration, il résultera un système sans fil hétérogène apportant de nouveaux défis dans la conception de l'architecture et des protocoles, le support de la mobilité et la QdS. Le système intégré devrait assurer une communication sans coupure avec une dégradation minimale de la QdS des usagers mobiles. Deux architectures génériques pour intégrer les réseaux WLAN et 3GPP/3GPP2 ont été

proposées dans la littérature dénommées respectivement *tight coupling* (couplage rigide) et *loose coupling* (couplage souple) (3GPP, 2004; 3GPP2, 2004). Une présentation sommaire des solutions basées sur ces modèles d'intégration est faite dans Lampropoulos *et al.* (2005). La Figure 2.1 montre les deux modèles d'intégration.

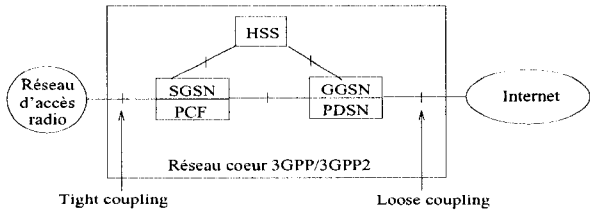

Figure 2.1 Architectures génériques d'intégration.

2.1.1 Exigences et réquis

La conception d'une architecture intégrée permettant une itinérance à travers des réseaux hétérogènes devrait respecter les exigences et requis suivants :

- *économique* : afin d'assurer un déploiement rapide et économique, une architecture intégrée doit utiliser l'infrastructure existante autant que possible et minimiser l'usage de nouvelles entités ;
- *évolutive* et *fiable* : l'intégration de n'importe quel nombre de systèmes appartenant à des opérateurs existants et futurs devrait être supportée ; de plus elle devrait être capable de garantir une tolérance aux pannes (*fault tolerance*) ;
- *signalisation* et *routage* : le trafic de signalisation ou de contrôle doit être maintenu en dessous d'un certain seuil ; de plus, le routage des données entre deux entités devrait être effectué à l'aide d'un chemin optimal ;
- *mobilité sans coupure* : afin d'éliminer les interruptions de connexion et la dégradation de la QdS durant une relève, l'architecture devrait supporter une mobilité sans coupure (*seamless mobility*) ;

- *sécurité* : l'architecture devrait fournir un niveau de sécurité qui serait équivalent sinon meilleur que celui offert par les réseaux existants.

Les exigences ci-dessus montrent qu'il est très difficile d'avoir une seule architecture qui serait appropriée pour tous les scénarios d'interopérabilité, et ainsi satisfaire tous les opérateurs et fournisseurs de services. Il est donc difficile de prédire quelle type d'architecture dominera le marché, car la sélection d'un modèle d'intégration n'est pas seulement basée sur des critères de performance, mais aussi sur les coûts et les profits. Ainsi, en attendant la conception et le déploiement d'une solution idéale, les usagers continueront d'exiger une solution pratique. Cela peut être obtenu grâce à un certain compromis sur les exigences et requis mentionnés précédemment. Plusieurs architectures de réseaux SFPG/4G ont été proposées ou sont en cours de définition (Kibria & Jamalipour, 2007). Cependant, beaucoup de questions demeurent ouvertes avant le déploiement des réseaux SFPG/4G.

2.1.2 Couplage rigide ou fort

Avec l'architecture *tight coupling* (couplage rigide), un réseau WLAN peut être considéré comme un réseau d'accès (*Radio Access Network* - RAN) complémentaire d'un réseau cellulaire 3G. En effet, dans cette approche le réseau WLAN est directement connecté au réseau cœur 3G de la même manière qu'un RAN de 3GPP/3GPP2. Ainsi, le trafic provenant du WLAN est acheminé avant tout à travers le réseau cœur du système cellulaire 3G avant d'atteindre un réseau de données externe tel que Internet. Dans cette approche, il est nécessaire de déployer une entité logique ou physique afin de permettre une transparence des caractéristiques du réseau WLAN et d'en rendre disponible certaines fonctionnalités du système 3G. Plusieurs protocoles disponibles dans les systèmes cellulaires 3G peuvent alors être réutilisés dans le réseau WLAN. La mobilité à travers les deux réseaux est

entièrement basée sur les protocoles de gestion de mobilité de 3GPP/3GPP2.

Le déploiement des réseaux cellulaires 3G se fait selon certaines règles de planification et d'ingénierie. La capacité et la configuration des éléments du réseau sont déterminées à partir des mécanismes spécifiques. Ainsi, une connexion directe du réseau WLAN au réseau 3G suscite des préoccupations aussi bien en terme de coûts que de capacité. En effet, injecter un trafic haut débit du réseau WLAN vers un réseau cellulaire 3G pourrait causer des problèmes à certains nœuds comme le SGSN (*Serving GPRS Support Node*) ou le GGSN (*Gateway GPRS Support Node*). Le fait que les interfaces du réseau cœur du système cellulaire 3G soient directement exposés au réseau WLAN, il serait plus approprié que les deux réseaux appartiennent au même opérateur. De plus, la carte réseau WLAN du terminal mobile devrait implémenter la pile de protocoles du réseau cellulaire 3G.

2.1.3 Couplage léger ou faible

Par contre, avec l'architecture *loose coupling* (couplage faible), le réseau WLAN est directement connecté au réseau de données externe sans passer par le réseau cœur du système cellulaire 3G tout en permettant aux usagers mobiles de bénéficier de services sans coupure. Il n'est donc pas nécessaire d'introduire des concepts du réseau 3G dans le réseau WLAN. L'approche *loose coupling* permet un déploiement et une ingénierie de trafic indépendants des réseaux WLAN et 3G sans un investissement majeur de capitaux. Aucun changement n'est nécessaire à l'architecture des réseaux et à la pile des protocoles utilisée. Cette approche utilise en grande partie les protocoles proposés par l'IETF, par exemple pour l'authentification, la facturation et la gestion de mobilité.

Toutefois, des exigences minimales sont requises au réseau WLAN et il peut être

nécessaire de déployer de nouveaux équipements pour la mise en correspondance de certaines spécifications. Pour garantir la mobilité sans coupure, les protocoles utilisés peuvent fonctionner indépendamment sur chaque réseau mais doivent être interopérables. Des accords d'itinérance entre opérateurs permettront aux usagers de bénéficier d'une couverture à plus grande échelle en faisant affaire avec un seul fournisseur de services ou opérateur. Cependant en terme de performance, on peut avoir des latences de relève et une perte de paquets considérables. Car chaque décision de relève nécessite de contacter le réseau nominal qui peut être assez loin de la position courante de l'abonné.

2.1.4 Couplage hybride

L'intérêt suscité par l'intégration et l'interopérabilite des réseaux WLAN et cellulaires 3G a conduit aussi à la proposition des approches hybrides. En effet, avec un couplage hybride, le chemin emprunté par les données est différencié selon le type de trafic (Song *et al.* , 2003). Le trafic temps-réel est achéminé en utilisant le couplage rigide tandis que le trafic non-temps réel utilise le couplage léger. On peut alors tirer profit des avantages de chacune des deux approches génériques. Toutefois, plusieurs inconvénients subsistent. Jaseemuddin (2003) propose une architecture intégrant les réseaux IEEE 802.11 et UMTS permettant à un nœud mobile de maintenir en parallèle une connexion pour les données à travers le réseau WLAN et pour la voix au moyen du réseau UMTS. Deux architectures pour intégrer les réseaux WLAN et cellulaire UMTS sont proposées par Salkintzis *et al.* (2002), basées sur les deux modèles génériques ci-dessus. Dans la première architecture (*tight coupling*), deux nouvelles entités appelées *GPRS Interworking Function* (GIF) et WAF (*WLAN Adaptation Function*) permettant l'interopérabilité des fonctionnalités entre les deux systèmes sont introduites.

Dans Buddhikot *et al.* (2003), une architecture intégrant les réseaux IEEE 802.11 et CDMA2000 basée sur l'approche *loose coupling* est proposée. Deux éléments caractérisent la proposition : un logiciel d'accès au service sur les terminaux des usagers (*client software*) et une passerelle appelée IOTA (*Integration Of Two Access technologies*) déployée dans le réseau WLAN. La passerelle IOTA permet le support de la mobilité inter-technologie, la garantie de la QdS et l'établissement des accords d'itinérance entre plusieurs fournisseurs de services et opérateurs. Dans Akyildiz *et al.* (2005), les auteurs proposent une architecture pour permettre l'interopérabilité des réseaux dans un environnement hétérogène. Deux nouvelles entités y sont introduites appelées NIA (*Network Interoperating Agent*) et IG (*Interworking Gateway*) pour permettre l'interaction des différentes technologies considérées et garantir une itinérance aux usagers. On peut à première vue prévoir la surcharge du NIA, même si les auteurs préconisent le contraire. La localisation de l'entité NIA générera sans aucun doute des problèmes de délai. En effet, si la relève inter-système a lieu dans un environnement où les différentes technologies se chevauchent (partiellement ou totalement) le délai de communication avec le NIA aura un impact négatif par exemple sur le trafic temps-réel. Aussi, la position du NIA aura un impact sur le routage des données, qui pourra être sous-optimal.

Le choix d'une architecture optimale d'intégration dépend de plusieurs facteurs. En effet, si par exemple un système sans fil hétérogène est composé d'un grand nombre de réseaux WLAN et 3G, l'architecture *loose coupling* est un choix approprié. D'autre part, si un opérateur est à la fois responsable des réseaux 3G et WLAN, l'architecture *tight coupling* est une option attractive. Dans le cas contraire, l'établissement des accords multilatéraux de mobilité entre les opérateurs est nécessaire. Toutefois, avec le nombre énorme d'opérateurs WLAN et 3G, cette tâche paraît très laborieuse. Bien qu'aucune conclusion ne soit totalement admise et acceptée, l'approche *loose coupling* offre plusieurs avantages au niveau architecture

avec des points faibles moins évidents (Buddhikot *et al.* , 2003) par rapport au modèle *tight coupling.* Ainsi, elle apparaît comme l'architecture préférée pour l'intégration des réseaux WLAN et 3G. Une architecture hybride permettant de tirer profit des avantages des deux approches devrait être proposée.

2.2 Mécanismes de mobilité

La gestion de mobilité a une influence considérable sur les performances des réseaux mobile et sans fil en particulier en ce qui concerne la qualité de service. Une classification des différents types de relèves dans les réseaux SFPG/4G est faite dans Nasser *et al.* (2006). Afin de gérer la mobilité dans les réseaux SFPG/4G, plusieurs protocoles ont été proposés dans la littérature en lien avec les différentes couches de la pile TCP/IP. Au niveau de la couche application, le protocole SIP (*Session Initiation Protocol*) a été proposé initialement pour gérer la signalisation pour des applications multimédia (Rosenberg *et al.* , 2002). Cependant, une extension a été apportée par la définition d'un nouveau message, appelé re-INVITE, qui permet de gérer la mobilité des usagers (Schulzrinne & Wedlund, 2000). Comme SIP utilise des protocoles de la couche transport pour acheminer ses messages de signalisation, il héritera des lacunes de ces protocoles dans un environnement sans fil et mobile. En outre, la mise à jour du serveur SIP dans le réseau nominal ainsi que le nœud correspondant après changement de localisation du nœud mobile, entraîne une augmentation du trafic de signalisation et de la latence de relève.

La gestion de mobilité au niveau de la couche transport a comme avantage d'éviter d'avoir une entité réseau servant de contrôleur et la transparence de la localisation du nœud mobile. Cependant, la couche transport est considérablement affectée par la mobilité des usagers. Il est donc nécessaire qu'un protocole de gestion de mobilité au niveau de la couche transport soit capable d'adapter rapidement

le trafic et les paramètres de congestion au nouveau réseau lors d'une relève. Le *Stream Control Transmission Protocol* (SCTP) est un nouveau protocole de la couche transport possédant les propriétés du *multi-streaming* et *multi-homing* (Ong & Yoakum, 2002). La propriété du *multi-homing* permet de maintenir plusieurs adresses IP lors d'une association, c'est-à-dire la connexion entre deux terminaux. Une extension de ce protocole, communément appelée *Mobile SCTP* (mSCTP) a été proposée par Koh *et al.* (2004, 2005) et Riegel & Tuexen (2006) afin d'assurer une interruption minimale d'une session en cours lors d'une relève. Les problèmes de signalisation, de latence et de perte de paquets subsistent avec mSCTP. De plus, un des problèmes majeurs avec mSCTP ou tout autre protocole utilisant une association directe entre un nœud mobile et son correspondant est la mobilité simultanée. Cette dernière réfère au contexte où les deux entités ayant une session en cours se déplace au même instant (Wong *et al.* , 2007).

La couche réseau ou IP apparaîssant *de facto* la plus appropriée pour assurer la convergence des différentes technologies d'accès, des protocoles de gestion de mobilité y ont aussi été proposés dans la littérature que l'on qualifie souvent de *IP-based Mobile Protocols.* En outre, avec la coexistence des diverses technologies d'accès dans les réseaux SFPG/4G, il est nécessaire de définr de nouveaux mécanismes de gestion de mobilité par exemple à travers la couche liaison. La gestion de mobilité abordée dans cette thèse portera sur ces deux dernières couches. Dans les sections qui suivent, nous allons effectuer un survol des différentes approches disponibles dans la littérature en lien avec nos objectifs de recherche.

2.2.1 Protocole Mobile IPv6

Le protocole *Mobile IPv6* (MIPv6) défini par l'IETF (Johnson *et al.* , 2004) est probablement le plus connu et sera le plus utilisé pour gérer la mobilité IP dans l'In-

ternet sans fil de prochaine génération. Sa simplicité et sa flexibilité lui ont permis d'être largement adopté au sein des organismes de standardisation et de la communauté scientifique. La relève dans MIPv6 peut être décrite comme une séquence des procédures suivantes : détection de mouvement, découverte de routeurs, configuration d'adresses, détection de la duplication d'adresses (*Duplicate Address Detection* - DAD), authentification et autorisation, enregistrement de l'adresse temporaire et mise à jour des associations. Le protocole MIPv6 permet aux nœuds mobile de maintenir une connexion au réseau lors de leurs déplacements et changements de point d'attache au réseau. Le nœud mobile (MN) dispose d'une adresse permanente (*Home Address* - HoA) associée à son réseau nominal (*Home Network*). Quand le MN se trouve dans un réseau visité (*Foreign Network*), il acquiert une adresse temporaire (*Care-of Address* - CoA) qui servira à son identification et au routage de ses données. La phase de découverte de réseau s'effectue à l'aide de l'échange des messages *Router Solicitation* (RS) et *Router Advertisement* (RA). Par contre, la procédure DAD est réalisée à partir des messages *Neighbor Solicitation* (NS) et *Neighbor Advertisement* (NA).

Chaque fois que le MN se déplace d'un réseau à un autre, il acquiert une nouvelle adresse CoA et envoie une requête de mise à jour (*Binding Update* - BU) de la cache d'association à son agent mère (*Home Agent* - HA) dans le réseau nominal afin d'effectuer la correspondance entre le CoA et le HoA. Le HA y répond par la transmission d'un message d'acquittement (*Binding Acknowledgment* - BAck). De la même manière, le MN effectue une mise à jour de la cache d'association auprès de tous ses correspondants (*Correspondent Node* - CN). L'acquisition d'adresses peut se faire de deux façons à savoir, l'autoconfiguration *stateful* et *stateless* (Thomson & Narten, 1998) et a lieu seulement en cas de relève au niveau de la couche IP. L'adresse temporaire CoA est utilisée comme adresse de routage tandis que l'adresse permanente HoA sert pour le transport et l'identification des applications.

Deux modes de routage entre le MN et son CN sont possibles : le routage triangulaire et le routage optimisé. Avec le routage optimisé, les paquets émis par le CN sont directement acheminés vers le MN sans passer par le HA. Par contre, avec le routage triangulaire, les paquets sont envoyés vers le HA, qui les interceptera avant de les transférer à l'adresse courante du MN. Pour supporter le routage optimisé, le MN doit entretenir l'association de sa cache auprès du CN. Avant la mise à jour de l'association au CN, la procédure *return routability* doit être effectuée afin de garantir l'authenticité du message BU, c'est-à-dire qu'il ne provient pas d'un nœud malicieux. Cette procédure est basée sur le test effectué sur l'adresse nominale (*home address test*) et un autre sur l'adresse temporaire (*care-of address test*). Dans le premier test, le MN envoie le message *Home Test Init* (HoTI) au HA qui le transfère au CN. Ce dernier y répond en envoyant le message *Home Test* (HoT) destiné à l'adresse nominale du MN en y incluant un jeton de sécurité (*Home Key Token*). Le HA transfère par la suite le message HoT vers l'adresse courante du MN. D'autre part, durant le test sur l'adresse temporaire, le MN envoie directement le message *Care-of Test Init* (CoTI) au CN, et ce dernier y répond à l'aide du message *Care-of Test* (CoT) contenant un jeton de sécurité (*Care-of Keygen Token*). La Figure 2.2 montre les opérations de base du protocole MIPv6.

Bien que le protocole MIPv6 ait été proposé pour supporter la mobilité au niveau de la couche IP, son analyse permet de déceler plusieurs faiblesses. Elles sont majeures en présence d'un taux élevé de mobilité des terminaux (relève rapide et fréquente). Parmi ces faiblesses, on peut citer : un délai (une latence) de relève et un taux de perte de paquets élevés, une mise à jour fréquente du HA et des CN – même suite à un petit déplacement du MN –, surcharge du trafic de signalisation, absence du support de la qualité de service, de la sécurité et de la radio-recherche (*paging*). La dégradation ou l'interruption d'une session en cours pourrait alors être observée. Cela est dû au fait que MIPv6 ne fait pas de distinction entre la

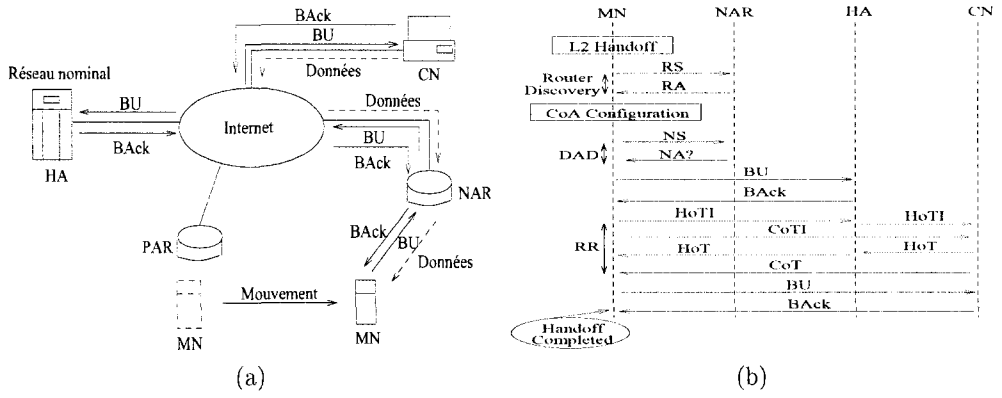

Figure 2.2 Protocole MIPv6 : (a) architecture; (b) signalisation.

micro-mobilité et la macro-mobilité. Il réagit de la même manière dans les deux cas. Ce qui est inacceptable pour des applications temps-réel et celles sensibles au débit. En fait, MIPv6 ne serait adapté qu'au trafic *best effort*. Une analyse de ces problèmes est faite dans Campbell *et al.* (2002), Chiussi *et al.* (2002), Reinbold & Bonaventure (2003) et Saha *et al.* (2004).

Afin de remédier aux faiblesses de MIPv6, d'autres protocoles ont été proposés servant ainsi de complément ou d'amélioration. Dans la plupart de ces protocoles, le support de la macro-mobilité est confié au protocole MIPv6, tandis que ces nouveaux mécanismes s'intéressent principalement à la micro-mobilité. L'approche de conception de ces protocoles semble parfois différer, mais leurs principes opérationnels sont largement similaires (Campbell *et al.* , 2002). En effet, ils introduisent soit une architecture hiérarchique permettant de localiser le trafic de signalisation et d'accélérer la mise à jour ou encore utilise une anticipation de la relève grâce aux informations de la couche liaison des données.

2.2.2 Hierarchical Mobile IPv6

Le protocole *Hierarchical Mobile IPv6* (HMIPv6) proposé par Soliman *et al.* (2005) a pour objectif de gérer la mobilité d'un nœud localement de façon à minimiser le trafic de signalisation dans le réseau et d'optimiser la performance lors de la relève. Il permet de réduire le délai de mise à jour de la localisation. Une nouvelle entité, appelée *Mobile Anchor Point* (MAP), est introduite pour assister les nœuds mobile lors d'une relève locale. Les mouvements des nœuds mobile à l'intérieur d'un domaine gérer par un MAP sont transparents pour les correspondants (CN) et l'agent nominal (HA). Avec HMIPv6, un MN est identifié à l'aide de deux adresses IP temporaires : une adresse RCoA (*Regional CoA*) pour le sous-domaine MAP et une adresse LCoA (*on-Link CoA*) qui correspond à sa localisation courante.

Si le nœud mobile (MN) change son adresse courante (LCoA) tout en demeurant à l'intérieur d'un domaine MAP (mouvement intra-MAP), il a uniquement besoin d'enregistrer cette nouvelle adresse auprès du MAP. Par contre, si le MN se déplace dans un autre domaine MAP (mouvement inter-MAP), il est nécessaire d'acquérir une nouvelle adresse RCoA et une nouvelle adresse LCoA, suivi de la mise à jour des associations au MAP et au HA/CN. Le MAP qui se comporte exactement comme un HA local interceptera tous les paquets destinés au MN qu'il dessert, puis décapsulera ces paquets pour enfin les transférer à l'adresse courante du MN. L'échange des messages de signalisation pour gérer la relève d'un MN ayant une session en cours est illustré à la Figure 2.3.

Avec HMIPv6, on observe encore une perte élevée des paquets (Vivaldi *et al.* , 2003), l'absence de la radio-recherche, des mécanismes de QdS et du support du trafic temps-réel. Une interruption de service pourrait donc se produire à cause du délai de relève et de la perte des paquets (Pérez-Costa *et al.* , 2003; Gwon *et al.* ,

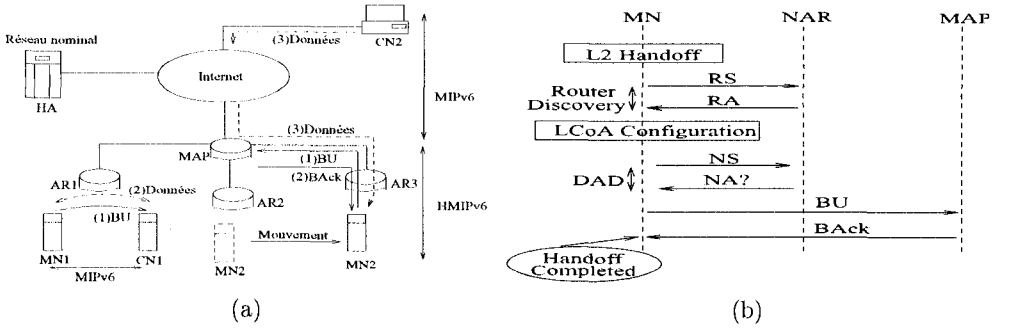

Figure 2.3 Protocole HMIPv6 : (a) architecture ; (b) signalisation.

2004). En outre, le MAP pourrait devenir un goulot ou un point de pannes potentiels, quoique une hiérarchisation à plusieurs niveaux est préconisée. Cependant, en présence d'une hiérarchie à plusieurs niveaux, le choix ou la sélection d'un MAP approprié pour un MN peut être délicat et aura un impact direct sur la QdS.

2.2.3 Fast Handovers for Mobile IPv6

Le protocole *Fast Handovers for Mobile IPv6* (FMIPv6) a été proposé par Koodli (2005) pour réduire le délai de relève du protocole MIPv6 et ainsi minimiser la période d'interruption de service durant la relève. Cette période d'interruption se produit à cause du temps requis pour que le MN fasse la mise à jour de son adresse CoA au HA après un mouvement entre deux routeurs d'accès (*Access Router* - AR). Le principe de FMIPv6 consiste à établir une nouvelle adresse temporaire avant que le MN ne rompe sa connexion avec l'ancien routeur d'accès. Avec les messages *Router Solicitation for Proxy* (RtSolPr) et *Proxy Router Advertisement* (PrRtAdv), FMIPv6 permet une anticipation du mouvement du MN. En effet, le MN peut détecter rapidement qu'il se déplace vers un nouveau réseau grâce à l'information pertinente fournie par la couche liaison (*L2 triggers*). Le but étant

de pouvoir réaliser la relève de la couche réseau (IP) avant que celle de la couche liaison ne soit terminée. Par conséquent, FMIPv6 permet de réduire le délai pour la détection de mouvement et celui de la configuration des adresses.

Un tunnel bi-directionnel est établi entre le précédent routeur d'accès (*Previous Access Router* - PAR) et le nouveau (*New Access Router* - NAR) suite à l'envoi d'un message *Fast Binding Update* (FBU) par le MN au PAR dont la réponse se fait à l'aide d'un message FBAck (*Fast Binding Acknowledgment*). Le message *Handover Initiate* (HI) envoyé par le PAR au NAR permet d'initier la relève. Il contient le PCoA (*Previous CoA*), l'adresse de la couche liaison et le NCoA (*New CoA*) du MN. En réponse au message HI, le NAR envoie un HAck (*Handover Acknowledgment*) au PAR. Afin de s'annoncer auprès du NAR, le MN envoie un message *Fast Neighbor Advertisement* (FNA). Le PAR interceptera et acheminera les paquets arrivant à l'adresse PCoA vers l'adresse NCoA tant que l'association du MN au NAR n'a pas encore été complétée. Une illustration des opérations du protocole FMIPv6 pour une relève initiée par le MN selon le mode prédictif est donnée à la Figure 2.4. Notons que FMIPv6 possède aussi un mode réactif. Ce dernier a lieu quand le message FBU n'a pas été envoyé via le PAR mais plutôt via le NAR. En effet, le message FBU est encapsulé dans le message FNA. Le NAR extrait le message FBU et le transfère au PAR. Ce dernier y répond en envoyant le message FBAck avant de commencer le transfert des paquets dont l'adresse de destination est le PCoA.

L'absence de synchronisation entre le temps que le MN se déplace vers le NAR et le début du transfert des paquets par le PAR, peut engendrer une perte de ceux-ci. En effet, une perte de paquets se produira si le transfert s'effectue trop rapidement ou trop tard par rapport au temps où le MN se détache du PAR et s'attache au NAR. Pour éviter ce problème, une technique de fenêtrage (*buffering*) peut être utilisée aux deux routeurs d'accès (PAR et NAR) ou encore le *bicasting*.

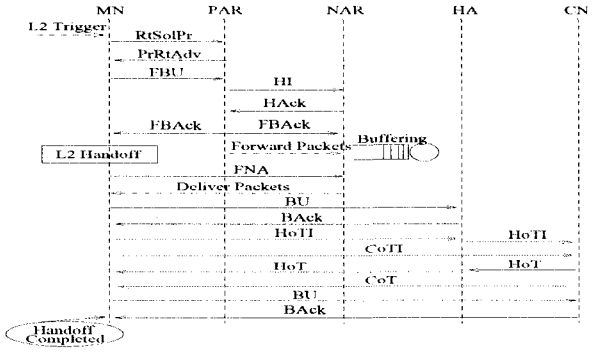

Figure 2.4 Opérations de base du protocole FMIPv6.

Avec le concept du *bicasting*, l'usager peut recevoir les paquets à partir des deux routeurs d'accès, ce qui permet de réduire leur perte. Cependant, le bicasting implique l'utilisation de deux liens de communication et l'assignation de deux adresses IP au MN, ce qui peut engendrer une duplication des paquets reçu et une allocation excessive de la bande passante sur la liaison sans fil. Il est donc nécesaire de bien définir l'intervalle de temps pour l'exécution du bicasting afin de réduire le nombre de paquets dupliqués et l'utilisation excessive de la bande passante. Une analyse de la perte de paquets avec FMIPv6 est faite par Kempf *et al.* (2003). Cette étude permet de se rendre compte que la perte des paquets est inévitable avec FMIPv6. En outre, la charge due à la signalisation avec FMIPv6 demeure importante. Alors, une dégradation du service subsiste. L'échec d'enregistrement à temps de l'adresse NCoA conduit à une performance de FMIPv6 similaire à celle de MIPv6. Une extension du protocole FMIPv6 consiste à différer l'obtention et l'utilisation de l'adresse CoA lorsqu'une session de trafic temps-réel est en cours.

2.2.4 Fast Handover for Hierarchical MIPv6

Les protocoles HMIPv6 et FMIPv6, développer chacun de leur côté ont pour objectif de réduire le trafic de signalisation et la latence de relève engendrés par le

protocole MIPv6. Afin de tirer profit des avantages de ces deux protocoles, il y a lieu d'examiner leur intégration. Toutefois, une simple intégration de FMIPv6 et HMIPv6 peut induire une surcharge de traitement non nécessaire (due à la redondance) pour le *tunneling* au routeur d'accès précédent et une utilisation inefficace de la bande passante (Jung *et al.* , 2005b). Une proposition d'intégration a été faite par Jung *et al.* (2005a) appelée *Fast Handover for Hierarchical Mobile IPv6* (F-HMIPv6). Dans F-HMIPv6, un tunnel, pour supporter une relève rapide, est établi entre le MAP et le NAR au lieu que ce soit entre le PAR et le NAR. Le MN échange les messages de signalisation FMIPv6 avec le MAP et non plus avec le PAR. Une illustration des opérations génériques du protocole F-HMIPv6 pour une relève initiée par le MN est donnée à la Figure 2.5.

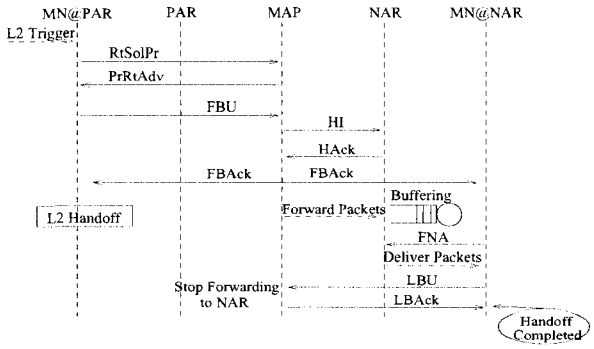

Figure 2.5 Opérations de base du protocole F-HMIPv6.

La perte de paquets dans F-HMIPv6 demeure et peut être supérieure à celle dans FMIPv6 et inférieure à celle dans HMIPv6 (Gwon *et al.* , 2004). En outre, puisque F-HMIPv6 résulte en une combinaison de HMIPv6 et FMIPv6, la charge de son trafic de signalisation sera plus élevée que celle de ces deux derniers protocoles pour une relève intra-MAP. D'autre part, avec F-HMIPv6 si le MN effectue une relève juste après avoir envoyé le message FBU au MAP, tous les paquets qui ont été transférés à l'ancienne adresse (*Previous on-Link CoA* - PLCoA), durant l'intervalle de temps nécessaire à la réception du FBU au MAP, seront perdus. Une

perte se produira aussi pour les paquets qui sont redirigés vers le MN lorsque la relève est initiée de la même manière que précédemment, ce qui augmenterait la latence de relève et le taux de perte de paquets. Une solution à ces problèmes est proposée par Pérez-Costa *et al.* (2003). Elle consiste à attendre aussi longtemps que possible (jusqu'à la perte de la connectivité) pour l'émission du message FBAck sur l'ancienne liaison pour commencer la relève. Une des faiblesses de F-HMIPv6, commune avec les autres protocoles de gestion de mobilité IP, est que le temps total d'interruption de service dépend du délai de la relève dans la couche liaison qui à son tour est fortement relié au nombre d'usagers dans le système.

2.2.5 Découverte de réseaux et transfert de contexte

Les protocoles de mobilité IP nécessitent *a priori* la connaissance du réseau cible avant de réaliser une relève en effectuant par exemple la découverte du prochain routeur d'accès. Cependant, cet aspect n'est pas souvent pris en compte de la description des protocoles ci-dessus. Afin d'avoir une mobilité sans coupure à travers différentes technologies d'accès et réseaux, un MN a besoin d'avoir l'information sur le prochain réseau vers lequel il devra se connecter. De plus, il est nécessaire de transférer les informations de sa session en cours (*context transfer*) du point d'attache actuel vers le prochain afin d'assurer la continuité de la connexion. Pour résoudre ces deux problèmes, l'IETF a proposé les protocoles *Candidate Access Router Discovery* (CARD) (Leibsch *et al.* , 2005) et *Context Transfer Protocol* (CXTP) (Loughney *et al.* , 2005). Ces deux protocoles permettent d'éviter l'usage inefficace des ressources d'un réseau sans fil qui sont limitées et assurent un transfert fiable du contexte. L'information transférée dans le contexte peut comprendre l'état de la QdS, les paramètres de sécurité et de facturation (AAA), l'état de la compression d'en-tête établi et maintenu entre le MN et le routeur d'accès.

L'objectif principal de ces deux protocoles est la réduction de la latence de relève et la perte de paquets, et d'éviter la ré-initiation de la signalisation entre le MN et le réseau à partir du début. Cependant, le transfert de contexte n'est pas toujours possible, par exemple quand le MN se déplace entre les réseaux d'administrateurs ou opérateurs différents. Le nouveau réseau peut requérir une re-authentification du MN et une initiation de la signalisation dès le début au lieu d'accepter l'information qui a été transférée. De plus, les entités échangeant le contexte doivent s'authentifier mutuellement. Cela peut être très laborieux à réaliser dans les réseaux SFPG/4G dû à la coexistence de technologies différentes et en particulier lors d'une relève verticale. D'autre part, il peut être nécessaire que le contexte de sécurité transféré soit adapté au nouvel environnement. Par exemple, l'adresse IP dans l'association de sécurité avec IPSec (Kent & Atkinson, 1998) peut changer ou encore différents mécanismes cryptographiques ou de protection du trafic pourraient être utilisés. De plus, les protocoles CARD et CXTP sont par nature réactif, donc ne peuvent pas permettre une collection dynamique des informations des routeurs d'accès voisins.

2.2.6 Autres protocoles de mobilité IP

Les protocoles FMIPv6 et HMIPv6 sont principalement orientés nœud mobile (*MN-based protocols*) similairement à MIPv6. Autrement dit, le MN a le contrôle de la gestion de mobilité. Cela nécessiterait un changement de la pile de logiciels sur le terminal. Toutefois, des problèmes de compatibilité peuvent apparaître avec les protocoles de gestion de mobilité globale de même qu'une complexité sur les procédures de sécurité. D'où le besoin d'un protocole de gestion de mobilité locale qui serait orienté réseau et ne nécessiterait pas une modification des logiciels au niveau du terminal (Kempf, 2007a,b).

Pour supporter la mobilité d'un nœud IPv6, une extension de la signalisation

de MIPv6 est faite de même qu'une réutilisation d'un HA via un agent de mobilité proxy (*proxy mobility agent*) dans le réseau. Le MN n'est plus impliqué dans la signalisation requise pour la gestion de mobilité. En effet, c'est l'agent de mobilité proxy qui effectue la signalisation pour gérer la mobilité au nom du MN. Le protocole *Proxy Mobile IPv6* (PMIPv6) (Gundavelli *et al.* , 2007), en cours de proposition à l'IETF, a pour objectif de fournir une gestion de mobilité orienté réseau (*network-based*) aux MNs, sans nécessiter la participation du MN dans l'échange des messages de signalisation lors de la mobilité.

Deux nouvelles entités fonctionnelles sont introduites dans l'architecture MIPv6. Le *Local Mobility Anchor* (LMA) qui est un HA pour le MN dans le domaine *Proxy Mobile IPv6.* Le *Mobile Access Gateway* (MAG) gère la signalisation nécessaire à la mobilité pour un MN qui s'attache à son lien d'accès. Par exemple, c'est lui qui s'occupe de l'enregistrement du MN auprès du LMA. Le MAG agit par défaut comme un AR et il dispose de l'information nécessaire pour émuler la liaison du réseau nominal du MN. Donc, il annonce le préfixe du réseau nominal du MN à ce dernier lui faisant ainsi croire qu'il est toujours sur le même lien. L'information contenu dans le message *Proxy Binding Acknowledgment* permet au MAG de connaître le préfixe du réseau nominal du MN. Durant ses déplacements dans domaine *Proxy Mobile IPv6*, MN a l'impression qu'il réside toujours sur le même que celui sur lequel il a obtenu son adresse initiale.

Dans Hsieh *et al.* (2003), les auteurs combinent les propositions faites dans Koodli (2005) et dans Soliman *et al.* (2005) pour concevoir une nouvelle architecture permettant de gérer une relève sans coupure de façon locale, appelée S-MIP (*Seamless Mobile IP*). Une nouvelle entité, appelée *Decision Engine* (DE), est ajoutée à l'architecture de HMIPv6 de même qu'une stratégie de synchronisation SPS (*Synchronized-Packet-Simulcast*). L'information sur les mouvements des usagers est contenue dans le DE ; il utilise la puissance du signal disponible à partir de la

couche liaison et l'identité des routeurs d'accès (AR). S-MIP permet de réduire le délai pour la détection de mouvements. Il fournit moins de pertes de paquets dues à la relève au niveau de la couche réseau au prix d'un accroissement minime du trafic de signalisation comparativement aux résultats obtenus par intégration de HMIPv6 et FMIPv6 décrits dans Soliman *et al.* (2005). S-MIP exploite l'hypothèse que les zones de couverture des différents routeur d'accès ne se chevauchent que partiellement. Ainsi, le protocole ne peut être étendu pour le support de la mobilité entre différents domaines, car la zone de couverture de l'un d'eux peut être complètement couverte par un autre dans un environnement sans fil hétérogène et hiérarchique (Akyildiz *et al.* , 2004).

Das *et al.* (2002) propose un protocole de gestion de mobilité intra-domaine appelé IDMP (*Intra-Domain Mobility Management Protocol*). Le protocole IDMP est basé sur une approche hiérarchique à deux niveaux afin de réduire la charge globale du trafic de signalisation et le délai de mise à jour. Deux nouvelles entités y sont définies soient, un *Mobility Agent* (MA) et un *Subnet Agent* (SA). Le MA est responsable de la gestion de la mobilité à l'intérieur d'un domaine tandis que le SA est chargé de la mobilité des nœuds dans un sous-réseau. Un nœud mobile dispose de deux adresses temporaires CoA : une adresse GCoA (*Global CoA*) qui spécifie le réseau auquel est attaché le nœud mobile et l'adresse LCoA (*Local CoA*) qui fournit la localisation du nœud mobile dans un sous-réseau. L'information de la couche L2 (*L2 triggers*) est utilisée afin d'avoir un mécanisme de relève rapide et minimiser ainsi la perte des paquets en cours de transmission (*in-flight packets*). Une extension de IDMP dans les réseaux mobile 4G permettant d'assurer la relève rapide et la radio-recherche (*paging*) est proposée dans Misra *et al.* (2002).

Bien que l'apport des différents protocoles de gestion de mobilité soit considérable et intéressant, on dénote un certain nombre de faiblesses. Ainsi, la proposition de nouveaux protocoles devrait tenir compte des avantages et inconvénients déjà

observés et aussi les exigences qui caractérisent les réseaux SFPG/4G. Il est difficile de trouver une solution pour la gestion de mobilité qui soit optimale pour n'importe quel type de réseaux et d'applications. Cela pourrait faire que plusieurs protocoles de gestion de mobilité puissent coexister.

2.2.7 Mécanismes de relève verticale

Dans les réseaux SFPG/4G une relève pourrait entraîner un changement du point d'attache au réseau au niveau de la couche liaison, ainsi qu'au niveau routage. Avec la coexistence des technologies d'accès distinctes, cette relève pourrait être verticale. En effet, avec l'intégration des sytèmes de communication distincts, un usager devra être capable d'être connecté au réseau lui offrant un meilleur service. Dans un environnement hétérogène, une relève peut être forcée ou volontaire. Dans un tel environnement, la sélection du réseau auquel un usager sera connecté est très important afin de lui garantir une meilleure QdS. Par exemple, 3GPP a approuvé la sélection basée sur l'identification du réseau d'accès (*Network Access Identifier* - NAI) (Ahmavaara *et al.* , 2003). Le NAI est identique à une adresse électronique contenant une portion pour identifier le nom de l'usager et une autre son réseau nominal (Aboba & Beadles, 1999) comme suit : username@domain.com.

Le défi majeur dans la sélection de réseau est de trouver le compromis le plus adéquat entre les préférences de l'usager, le type d'application et les conditions du réseau. Le processus de sélection de réseau peut être subdivisé en trois étapes. La première consiste en une collecte des informations nécessaires telles que, le type d'application, les préférences de l'usager et les conditions du réseau qui auront un impact sur la décision finale. La deuxième étape consiste à utiliser les informations collectées comme données d'un algorithme de gestion de relève qui a pour objectif d'assurer une meilleure connectivité (*always best connected*) à l'usager (Gustaf-

sson & Jonsson, 2003). Autrement dit, l'usager ne bénéficie pas seulement d'une connectivité mais aussi une meilleure QdS en tout temps et n'importe où. La dernière phase correspond à la prise de décision par rapport aux résultats obtenus avec l'algorithme de gestion de relève.

Afin de gérer une relève verticale, plusieurs approches ont été proposées dans la littérature, chacune ayant des avantages et inconvénients. De façon classique, la décision de relève est basée sur la qualité du signal reçu (*received signal strength - RSS*) et la disponibilité du canal de communication. Cependant, une comparaison directe de la puissance du signal reçu à partir de deux technologies distinctes n'est pas possible ou peut entraîner une mauvaise interprétation. Ainsi, cette comparaison n'est pas suffisante pour gérer la relève dans un environnement hétérogène. Il est donc nécessaire de prendre en compte en plus de ces deux paramètres d'autres facteurs tels que, le coût monétaire, les conditions du réseau, le taux de transmission des données, les préférences de l'usager et la sécurité (McNair & Zhu, 2004).

Une décision de relève basée sur tous ces paramètres est cruciale dans les réseaux SFPG/4G, mais demeure cependant une question ouverture. En effet, certains de ces facteurs sont difficiles à quantifier. Ils peuvent être décrits comme suit :

- *Coût monétaire* : le coût ou prix est une considération primordiale pour les usagers car différents opérateurs peuvent avoir différentes stratégies de facturation. Cette différence de facturation peut affecter le choix des usagers lors d'une relève.
- *Énergie de la batterie* : pour certaines relèves, l'énergie de la batterie peut être un facteur significatif. Par exemple, quand la batterie est faible, l'usager pourrait décider de se connecter vers un réseau n'ayant pas des exigences trop élevées en termes de puissance.
- *Conditions du réseau* : les paramètres réseau tels que le trafic, la bande pas-

sante disponible, le délai, la congestion et la perte des paquets peuvent être considérés pour une utilisation efficace du réseau. La prise en compte de l'information du réseau pourrait être utile pour assurer une meilleure répartition de charges parmi les différents réseaux, permettant ainsi de réduire la congestion dans certains systèmes.

- *Performance du système* : pour garantir une performance adéquate du système, plusieurs paramètres tels que les caractéristiques du canal radio, l'affaiblissement de parcours, l'interférence inter-canal, le rapport signal à bruit (*signal to noise ratio - SNR*) et le taux d'erreurs sur les bits (*bit error rate* - BER) peuvent être utilisés pour la décision de relève.
- *Types d'application* : différents types d'application exigent des niveaux distincts de fiabilité, de délai et de débit. Les applications en cours sur un terminal mobile peuvent aussi influencer la décision de relève.
- *Conditions du nœud mobile* : les conditions du nœud mobile incluent les facteurs dynamiques tels que la vitesse, les informations de localisation et le modèle de mobilité (*moving pattern*).
- *Préférences de l'usager* : les préférences de l'usager peuvent être utilisées pour des requêtes spéciales pour un système par rapport à un autre.

Une architecture permettant d'intégrer un réseau 3GPP2 au WLAN est proposée dans Buddhikot *et al.* (2003) de même qu'un mécanisme pour la sélection d'interfaces radio. Cette sélection est basée sur la puissance du signal et sur la priorité attribuée à chaque interface. Tel que mentionné précédemment, ces paramètres ne sont pas suffisants pour une décision de relève dans les réseaux SFPG. En outre, dans le mécanisme proposé par Buddhikot *et al.* (2003), le MN doit évaluer continuellement de manière passive les conditions d'un besoin de relève même si l'application ou la session en cours bénéficie d'une meilleure qualité de service à travers le réseau qui dessert l'usager. Ceci engendre une consommation inutile des

resources du réseau et une utilisation excessive de la batterie du terminal.

Afin de réduire la consommation de l'énergie de la batterie d'un terminal mobile, sans dégrader le niveau du débit, une approche appelée *Wise Interface SElection* (WISE) a été proposée pour gérer les relèves verticales entre les réseaux 3G et WLAN (Minji *et al.* , 2004). WISE introduit une nouvelle entité appelée *Virtual Domain Controller* (VDC) dans le réseau cœur 3G, agissant comme un point de contrôle centralisé entre les deux technologies. La décision de relève avec WISE est effectuée en prenant en compte la charge du réseau et l'énergie consommée sur chaque interface radio. Cependant, le VDC constitue un point potentiel de pannes pour une telle architecture. En outre, les règles de décision de relève sont étroitement liées à la qualité du signal et à la bande passante disponible. Une approche utilisant le concept du *IP-based mobile protocol* est proposée par Du *et al.* (2002) pour gérer les relèves aussi bien verticales qu'horizontales, denommée HOPOVER (*HandOff Protocol for OVERlay networks*). Bien que HOPOVER permette une réduction du trafic de signalisation, il requiert un maintien excessif des informations sur les nœuds mobile auprès des points d'accès. De plus, pour permettre l'échange des messages de signalisation entre différentes entités, la procédure de relève de HOPOVER nécessite une standardisation complète entre les différents opérateurs.

Pour assurer un meilleur contrôle et une meilleure gestion des ressources réseau, afin de garantir une QdS appropriée, une architecture basée sur les politiques a été proposée par l'IETF (Yavatkar *et al.* , 2000). Cette architecture a motivé le travail décrit dans Wang *et al.* (1999) où une fonction de coût est proposée pour modéliser la décision de relève en prenant en compte plusieurs facteurs tels que la puissance, la bande passante et le prix des communications. Cependant, la fonction de coût proposée est très primaire et ne peut pas être utilisée pour des scénarios plus complexes. Afin de maximiser la QdS perçue par les usagers, deux algorithmes de décision pour les relèves verticales sont décrits dans McNair & Zhu (2004). Cepen-

dant, le problème d'une relève instable n'est pas examiné, de même que la gestion de mobilité au niveau IP, autrement dit l'impact que cette décision aura sur une relève au niveau IP. En outre, la fonction de décision proposée peut entraîner des problèmes de singularités, si par exemple il n'y a aucun frais pour les connexions. Dans Song & Jamalipour (2005), un mécanisme de sélection de réseau dans un environnement intégré 3G/WLAN est proposé afin de fournir une QdS adéquate aux usagers en tout temps. Cependant, la complexité opérationnelle de l'algorithme de décision introduit certains problèmes d'implémentation et empêche son application en temps-réel dans un environnement très dynamique. Un survol d'autres mécanismes de décision de relève est disponible dans Zhu & McNair (2006).

Un des points faibles communs aux mécanismes de gestion de relèves verticales disponibles dans la littérature est lié à l'inefficacité de la gestion des interfaces radio. En effet, une activation continuelle ou périodique des interfaces est souvent utilisée. Ces deux approches entraînent une consommation excessive de l'énergie des batteries et des ressources réseau. Il est donc nécessaire d'avoir des mécanismes plus efficaces et intelligents de gestion d'interfaces. D'autre part, dans un environnement hétérogène multi-accès, un MN doit être capable de détecter de façon aisée et efficace la présence et la disponibilité d'un nouveau réseau. La question sous-jacente est de savoir, quand et comment activer les interfaces radio qui sont en veille (*idle interface*) afin de détecter les autres technologies ou réseau d'accès sur la zone de couverture. Le compromis entre l'efficacité énergétique et le délai de découverte des réseaux d'accès dépend beaucoup de la solution à cette question.

CHAPITRE 3

AN ANALYTICAL FRAMEWORK FOR PERFORMANCE EVALUATION OF IPV6-BASED MOBILITY MANAGEMENT PROTOCOLS

Christian Makaya and Samuel Pierre
Mobile Computing and Networking Research Laboratory (LARIM)
Department of Computer Engineering, École Polytechnique de Montréal
P.O. Box 6079, Station Centre-ville, Montréal, Québec, H3C 3A7, Canada
Email : {christian.makaya, samuel.pierre}@polymtl.ca

Abstract

Mobility management with provision of seamless handover is crucial for an efficient support of global roaming of mobile nodes (MNs) in next-generation wireless networks (NGWN). Mobile IPv6 (MIPv6) and its extensions were proposed by the IETF for IP layer mobility management. However, performance of IPv6-based mobility management schemes is highly dependent on traffic characteristics and user mobility models. Consequently, it is important to assess this performance in-depth through those two factors. The performance of IPv6-based mobility management schemes is usually evaluated through simulations. This paper proposes an analytical framework to evaluate the performance of IPv6-based mobility management protocols. This proposal does not aim to advocate which is better but rather to study the effects of various network parameters on the performance of these protocols to enlighten decision-making. The effect of system parameters, such as subnet residence time, packet arrival rate and wireless link delay, is investigated for per-

formance evaluation with respect to various metrics like signaling overhead cost, handoff latency and packet loss. Numerical results show that there is a trade-off between performance metrics and network parameters.

Keywords : Analytical modeling, IP mobility protocols, mobility management, performance evaluation, quality of service, wireless networks.

3.1 Introduction

Next-generation wireless networks (NGWN) or fourth generation wireless networks (4G) are expected to exhibit heterogeneity in terms of wireless access technologies and services. With NGWN/4G, mobile nodes (MNs) or subscribers will have more demands for seamless roaming across different wireless networks, support of various services (e.g., multimedia applications) and quality of service (QoS) guarantees. Conceptually, the NGWN architecture can be viewed as many overlapping wireless access domains (e.g., UMTS, CDMA2000, WLAN, WiMAX). However, this heterogeneity brings new challenges for architecture design, mobility management, QoS provision and security. Moreover, heterogeneity in terms of radio access technologies and network protocols in NGWN requires common interconnection elements. Since the Internet Protocol (IP) technology enables the support of applications in a cost effective and scalable way, it is expected to become the core or backbone network of NGWN (Akyildiz *et al.* , 2005). Thus, current trends in communication networks evolution are directed towards all-IP principles in order to hide the heterogeneity and achieve convergence of these various networks.

Mobility management with provision of seamless handoff is key topic in NGWN. Then, it is crucial to provide seamless mobility and service continuity in intelligent and efficient ways. The Internet Engineering Task Force (IETF) has proposed

Mobile IPv6 or MIPv6 (Johnson *et al.* , 2004) as the main protocol for mobility management at the IP layer. However, MIPv6 has some well known drawbacks such as signaling traffic overhead, especially when the home agent (HA) or the correspondent node (CN) is located geographically far away from the mobile node (MN). Message transmission time for binding update registration will become very high resulting in long delay (handoff latency) and high packet loss rate thereby causing user-perceptible deterioration of real-time traffic.

Then, several extensions such as Fast Handovers for MIPv6 (FMIPv6) (Koodli, 2005) and Hierarchical MIPv6 (HMIPv6) (Soliman *et al.* , 2005), have been proposed to enhance the performances of MIPv6. In spite of these extensions, mobility management with QoS provision in NGWN remains a challenging and complex task. Usually, performance evaluation of IP-based mobility management schemes is based on simulation and testbed approaches and most available work focuses on these aspects (Pérez-Costa *et al.* , 2003; Gwon *et al.* , 2004). However, scenarios used for simulations vary greatly, the comparison of IP-based handoff protocols is hardly possible. Few works are available in the literature which assess IPv6-based mobility management protocols through analytical models. On the other hand, they are often based on simple assumptions and have some drawbacks.

In Xie & Akyildiz (2002), trade-off relationship between location update cost and packet tunneling cost is introduced in order to compute total signaling cost and evaluate the efficiency of IP-based mobility protocols. Work presented in Xie & Akyildiz (2002) is largely based on concepts introduced for location management in personal communication systems (PCS). Analytical models for handoff latency of IPv6-based mobility protocols are presented in Pérez-Costa *et al.* (2002) in order to assess the most appropriate scheme for functional specification and implementation. Analysis of signaling bandwidth according to binding update emission frequency is presented in Castelluccia (2000). However, signaling overhead generated by packets

tunneling is not considered. An analytical model for performance evaluation of HMIPv6 in IP-based cellular networks was proposed in Pack & Choi (2003). This model ignores periodic binding refresh and binding lifetime period, which may significantly affect total signaling cost. Moreover, the packet delivery cost only takes bandwidth consumption into account for data and ignores the extra signaling consumption due to control traffic. An analysis of the FMIPv6 signaling overhead is compared to that of MIPv6 in Pack & Choi (2004). However, packet loss, handoff latency and the impact of user mobility models were not investigated.

Contrary to previous works, in this paper, we perform a comprehensive analysis of various IPv6-based mobility protocols proposed by the IETF. We derive signaling traffic overhead, packet delivery, binding refresh and total signaling costs generated by an MN during its subnet residence time for each protocol. Moreover, the required buffer space, handoff latency and packet loss expressions are derived. The effect of mobility and traffic parameters on these criteria are analyzed from numerical results. The remainder of this paper is organized as follows : the next section offers a brief overview of IPv6-based mobility management schemes. After that, the proposed analytical framework is presented. Numerical results based on this analytical model is then investigated before concluding remarks drawn in the last section.

3.2 IP-based Mobility Management Protocols

Mobility management enables systems to locate roaming users in order to deliver data packets, i.e., *location management* and maintain connections with them when moving into a new subnet, i.e., *handover management*. Several protocols have been proposed for these purposes for IP mobility and are briefly presented in this section.

Definition : A *handover* or *handoff* is a movement of an MN between two attachment points, i.e., the process of terminating existing connectivity and obtaining new connectivity. Handovers in IP-based NGWN may involve changes of the access point at the link layer and routing path changes at the IP layer.

Definition : The *handoff latency* at an MN side is the time interval during which an MN cannot send or receive any packets during handoff and it is composed of L2 (link layer) and L3 (IP layer) handoff latencies. The L3 handoff latency is the sum of delay due to : movement detection, IP addresses configuration and binding update procedure.

Definition : The *signaling traffic overhead* is defined as the total number of control packets exchanged between an MN and a mobility agent (e.g., home agent). Efficient mechanisms must ensure seamless handover, i.e., with minimal signaling overhead, handoff latency, packet loss, and handoff failure and services continuity.

3.2.1 Mobile IPv6 (MIPv6)

MIPv6 was proposed for mobility management at the IP layer and allows an MN to remain reachable despite its movement within the IP environment. Each MN is always identified by its home address (HoA). While away from its home network, an MN is also associated with a care-of address (CoA), which provides information about the MN's current location. Discovery of new access router (NAR) is performed through Router Solicitation/Advertisement (RS/RA) messages exchange. Furthermore, to ensure that a configured CoA, through stateless or stateful mode (Thomson & Narten, 1998), is likely to be unique on the new link, the Duplicate Address Detection (DAD) procedure is performed by exchanging Neighbor Solici-

tation/Advertisement (NS/NA) messages. After acquiring a CoA, an MN performs binding update to the home agent (HA) through binding update (BU) and binding acknowledgment (BAck) messages exchange. To enable route optimization, BU procedure is also performed to all active CNs.

However, return routability (RR) procedure must be performed before executing a binding update process at CN in order to insure that BU message is authentic and does not originate from a malicious MN. The return routability procedure is based on home address test, i.e., Home Test Init (HoTI) and Home Test (HoT) messages exchange, and care-of address test, i.e., exchange of Care-of Test Init (CoTI) and Care-of Test (CoT) messages. Although RR procedure helps to avoid session hijacking, it increases delay of the BU procedure. Fig. 3.1(a) represents the sequence of message flow used in MIPv6 based on stateless address autoconfiguration.

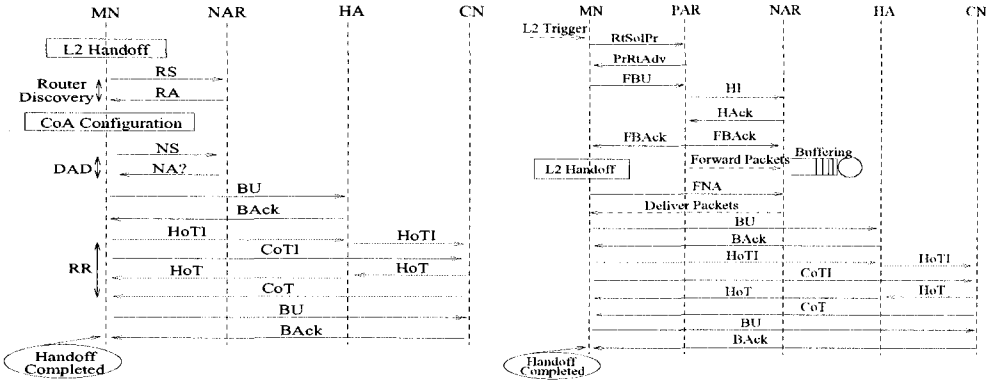

Figure 3.1 Signaling messages sequence : (a) MIPv6; (b) FMIPv6.

Analysis of MIPv6 shows that it has some well-known disadvantages such as overhead of signaling traffic, high packet loss rate and handoff latency, thereby causing user-perceptible deterioration of real-time traffic. Furthermore, the scala-

bility problems arise with MIPv6 since it handles MN local mobility in the same way as global mobility. Simultaneous mobility is another problem MIPv6 faces due to route optimization, which can occur when two communicating MNs have ongoing session and they both move simultaneously (Wong *et al.* , 2007). These weaknesses have led to the investigation of other solutions to enhance MIPv6 performance.

3.2.2 Fast Handovers for Mobile IPv6 (FMIPv6)

FMIPv6 was proposed to reduce handoff latency and minimize service disruption during handovers pertaining to MIPv6. The link layer information (L2 trigger) is used either to predict or rapidly respond to handover events. When an MN detects its movement toward NAR, by using L2 trigger, it exchanges Router Solicitation for Proxy (RtSolPr) and Proxy Router Advertisement (PrRtAdv) messages with the previous access router (PAR) in order to obtain information about NAR and to configure a new CoA (NCoA). Then, the MN sends a Fast Binding Update (FBU) to PAR in order to associate previous CoA (PCoA) with NCoA. A bi-directional tunnel between PAR and NAR is established to prevent routing failure with Handover Initiate (HI) and Handover Acknowledgment (HAck) message exchanges.

The Fast Binding Acknowledgment (FBAck) message is used to report status about validation of pre-configured NCoA and tunnel establishment to MN. Moreover, the PAR establishes a binding between PCoA and NCoA and tunnels any packets addressed to PCoA towards NCoA through NAR's link. The NAR buffers these forwarded packets until the MN attaches to NAR's link. The MN announces its presence on the new link by sending Router Solicitation (RS) message with the Fast Neighbor Advertisement (FNA) option to NAR. Then, NAR delivers the buffered packets to the MN. The sequence of messages used in FMIPv6 is illustrated in Fig. 3.1(b) for MN-initiated handoff of predictive mode.

A counterpart to predictive mode of FMIPv6 is reactive mode. This mode refers to the case where the MN does not receive the FBack on the previous link since either the MN did not send the FBU or the MN has left the link after sending the FBU (which itself may be lost), but before receiving a FBack. In the latter case, since an MN cannot ascertain whether PAR has successfully processed the FBU, it forwards a FBU, encapsulated in the FNA, as soon as it attaches to NAR. If NAR detects that NCoA is in use (address collision) when processing the FNA, it must discard the inner FBU packet and send a Router Advertisement (RA) message with the Neighbor Advertisement Acknowledge (NAACK) option in which NAR may include an alternate IP address for the MN to use. Otherwise, NAR forwards FBU to PAR which responds with FBack. At this time, PAR can start tunneling any packets addressed to PCoA towards NCoA through NAR's link. Then, NAR delivers these packets to the MN.

3.2.3 Hierarchical Mobile IPv6 (HMIPv6)

With MIPv6, an MN performs binding update to HA/CNs regardless of its movements to other subnets. This induces unnecessary signaling overhead and latency. To address this problem, HMIPv6 was proposed to handle handoff locally through a special node called Mobility Anchor Point (MAP). The MAP, acting as a local HA in the visited network, will limit the amount of MIPv6 signaling outside its domain and reduce the location update delay. An MN residing in a MAP's domain is configured with two temporary IP addresses : a regional care-of address (RCoA) on the MAP's subnet and an on-link care-of address (LCoA) that corresponds to the current location of the MN.

As long as an MN moves within MAP's domain or access network (AN) it does not need to transmit BU messages to HA/CNs, but only to MAP when its LCoA

changes. Hence, the movement of an MN within MAP domain is hidden from HA/CNs. For inter-MAP domain roaming, MIPv6 is used rather than HMIPv6. When an MN crosses a new MAP's domain, moreover from registering with new MAP, BU messages need to be sent by the MN to its HA/CNs to notify them of its new virtual location. Fig. 3.2(a) presents the generic sequence of message flows used in HMIPv6 with assumption that an MN has entered into new MAP domain and MIPv6 registration procedure was already completed.

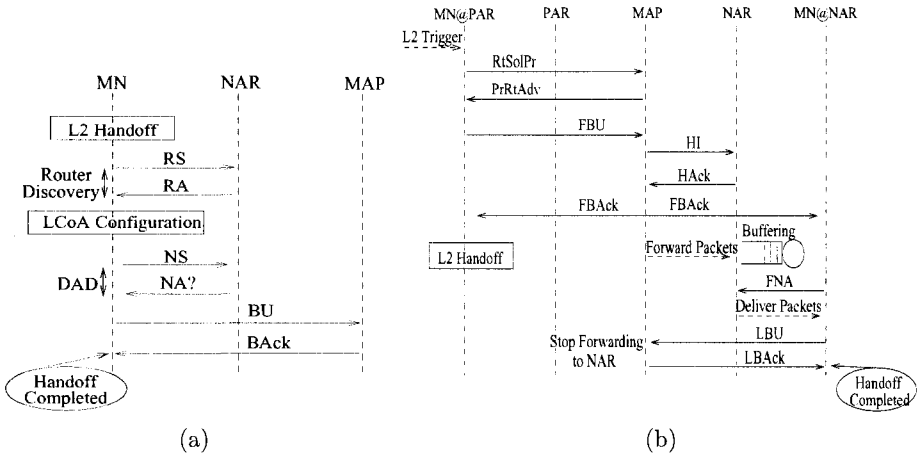

Figure 3.2 Signaling messages sequence : (a) HMIPv6; (b) F-HMIPv6.

3.2.4 Fast Handover for HMIPv6 (F-HMIPv6)

Combination of HMIPv6 and FMIPv6 motivates the design of Fast Handover for Hierarchical Mobile IPv6 (F-HMIPv6) protocol (Jung *et al.* , 2005a) in order to allow more efficient network bandwidth usage similarly to HMIPv6. Furthermore, like FMIPv6, it aims to reduce the handoff latency and packet loss. In F-HMIPv6, the bi-directional tunnel is established between MAP and NAR, rather than between PAR and NAR as it is in FMIPv6. After signaling message exchanges (between an

MN and the MAP) based on FMIPv6 messages, an MN follows the normal HMIPv6 operations by sending local BU (LBU) to MAP. When MAP receives LBU with the new LCoA (NLCoA) from MN, it will stop packets forwarding to NAR and then clear the established tunnel.

In response to LBU, the MAP sends local BAck (LBAck) to the MN and the remaining procedure follows the operations of HMIPv6. In the original F-HMIPv6 proposal, when handover anticipation cannot be supported, regular operations of HMIPv6 are used (Jung *et al.* , 2005a). Hence, HMIPv6 corresponds to reactive mode of F-HMIPv6. Fig. 3.2(b) illustrates a sequence of message used in F-HMIPv6 when an MN moves from PAR to NAR within MAP's domain and the MAP already knows the adequate information on the link-layer address and network prefix of each AR. This illustration is based on the assumption that an MN has entered into a new MAP domain and that MIPv6/HMIPv6 registration procedures were already completed.

3.3 Analytical Models

In IPv6-based wireless networks, QoS may be defined by packet loss, handoff latency and signaling traffic overhead. Analysis of these metrics is very useful to assess the performance of mobility management protocols in IP-based mobile environments. An analytical framework for evaluating performance of IP mobility protocols is proposed in this section. The notation used in this paper is given in Table 3.1.

Let χ_T be the random variable for the time between L2 trigger generation and link down (i.e., pending L2 handover) and $f_T(u, \sigma)$ the probability density function for successful completion of signaling, where $\sigma > 0$ is a success rate parameter. The

Tableau 3.1 Notation.

t_c	subnet (AR's coverage area) residence time random variable
t_d	AN/MAP domain residence time random variable
f_c(resp. f_d)	probability density function (PDF) of t_c (respectively t_d)
t_s	inter-session time between two consecutive sessions with PDF f_s
N_c	number of subnets crossing during intra-AN/MAP handoffs
N_d	number of AN/MAP domain crossing during inter-AN/MAP handoffs
C^g	global binding update cost to HA/CNs
C^l	local binding update cost to MAP
M	number of subnets in AN/MAP domain
N_{CN}	number of CNs having a binding cache entry for an MN
$d_{X,Y}$	average number of hops between nodes X and Y
$C_{X,Y}$	transmission cost of control packets between nodes X and Y
PC_X	processing cost of control packet at node X
C_{hc}	binding update cost at HA and CNs
C_{rr}	signaling cost for return routability procedure
t_T	time period from the L2 trigger to the starting of link switching

probability P_s of anticipated handover signaling success for a particular observed valued t_T is expressed as follows :

$$P_s = Pr(\chi_T > t_T) = \int_{t_T}^{\infty} f_T(u, \sigma)du. \tag{3.1}$$

Deriving an expression of P_s is difficult, as it depends on the exact form of $f_T(u, \sigma)$, which is usually unknown. For the sake of simplicity, we assume that χ_T is exponentially distributed.

3.3.1 User Mobility and Traffic Models

User mobility and traffic models are crucial for efficient system design and performance evaluation. We consider a traffic model composed of two levels, a session and packet. Usually, MN mobility is modeled by the cell residence time and various

types of random variables are used for this purpose (Fang, 2003). In NGWN, although the incoming calls or sessions follow the Poisson process (i.e., inter-arrival time are exponentially distributed), the inter-session arrival times may not be exponentially distributed (Fang, 2003). Other distribution models, like hyper-Erlang, Gamma and Pareto have been proposed to model various time variables in wireless networks. However, performance evaluations reported in the literature (Fang, 2003) show that exponential model can be appropriate for cost analysis. In fact, exponential model provides an acceptable trade-off between complexity and accuracy.

Let μ_c and μ_d be the border crossing rate of an MN out of a subnet (AR) and out of an access network (AN) or MAP domain, respectively. Furthermore, let μ_l be the border crossing rate for which the MN still stays in the same AN/MAP domain. When an MN crosses an AN/MAP domain border, it also crosses an AR border. Then, according to Baumann & Niemegeers (1994), if we assume that the AN/MAP coverage area is circular with M subnets each with size a_{AR}, the border crossing rates are given by :

$$\mu_d = \frac{\mu_c}{\sqrt{M}} \quad \text{and} \quad \mu_l = \mu_c - \mu_d = \mu_c \frac{\sqrt{M}-1}{\sqrt{M}} \tag{3.2}$$

where $\mu_c = 2\frac{v}{\sqrt{\pi a_{AR}}}$, v is the average velocity of an MN, $a_{AR} = \pi R^2$ and R is the radius of access router coverage area or subnet.

Modeling the probability distribution of the number of boundary crossing during a call plays a significant role in cost analysis for wireless cellular networks. This will be the case again for IP-based wireless networks. Fig. 3.3 shows the timing diagram for typical mobile user crossing access router i (AR_i) boundary and moving to AR_j during inter-session time. t_{rs} denotes a residual subnet residence time. In case of inter-AN/MAP movement, a similar figure for timing diagram for access network boundary crossing may be drawn by replacing AR by MAP, t_c by t_d and t_{rs} by the

residual access network residence time (t_{ra}).

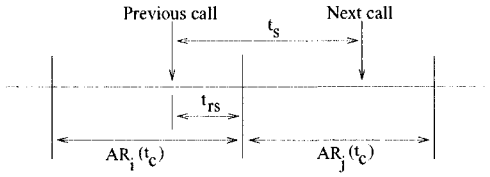

Figure 3.3 Timing diagram for subnet boundary crossing.

According to the notation of Table 3.1 and the timing diagram illustration, the subnet crossing probability (P_c) and AN/MAP domain crossing probability (P_d) during inter-session time interval are expressed as follows :

$$\begin{aligned} P_c &= Pr(t_s > t_c) = \int_0^\infty Pr(t_s > u) f_c(u) du \\ P_d &= Pr(t_s > t_d) = \int_0^\infty Pr(t_s > u) f_d(u) du. \end{aligned} \tag{3.3}$$

The probability that an MN experiences k subnets boundary crossings and m access network boundary crossings during its session lifetime corresponds to probability mass function of N_c and N_d, respectively and expressed as follows (Xiao *et al.* , 2004) :

$$Pr(N_c = k) = P_c^k(1 - P_c) \quad \text{and} \quad Pr(N_d = m) = P_d^m(1 - P_d). \tag{3.4}$$

Then, the average number of location binding updates during an inter-session time interval under subnet crossing ($E(N_c)$) and AN/MAP domain crossing ($E(N_d)$) are given by :

$$\begin{aligned} E(N_c) &= \sum_{k=0}^{\infty} kPr(N_c = k) = \sum_{k=0}^{\infty} kP_c^k(1 - P_c) \\ E(N_d) &= \sum_{m=0}^{\infty} mPr(N_d = m) = \sum_{m=0}^{\infty} mP_d^m(1 - P_d). \end{aligned} \tag{3.5}$$

For simplicity and easy derivation of signaling cost, exponential assumption is

made. In other words, we assume that residence time in a subnet and in AN/MAP domain follow exponential distribution with parameters μ_c and μ_d, respectively while session arrival process follows a Poisson distribution with rate λ_s. Hence, boundary crossing probabilities and average number of location updates during an inter-session time interval can be easily obtained as follows :

$$\begin{aligned} P_c &= \frac{\mu_c}{\mu_c + \lambda_s} \quad \text{and} \quad P_d = \frac{\mu_d}{\mu_d + \lambda_s} \\ E(N_c) &= \frac{\mu_c}{\lambda_s} \quad \text{and} \quad E(N_d) = \frac{\mu_d}{\lambda_s}. \end{aligned} \tag{3.6}$$

Similarly, we can derive the expression of the average number of subnets, $E(N_l)$, that an MN crosses but still stay within AN/MAP domain during an inter-session time interval.

3.3.2 Total Signaling Cost

Performance analysis of wireless networks should consider a total signaling cost induced by mobility management schemes. As for wireless cellular networks, signaling traffic overhead cost must be evaluated for NGWN or IP-based mobile environments. In NGWN, there are two kinds of location update signaling. One occurs from an MN's subnet crossing and the other occurs when the binding is about to expire. To differentiate them, the former refers to binding update (BU) message and the last one refers to binding refresh (BR) message. Moreover, delivery of data packets induces usage of network resources, then generates an additional cost. Thus, the total signaling cost, C_T, could be considered as the sum of binding update signaling cost, C_{BU}, binding refresh signaling cost, C_{BR}, and packet delivery cost, C_{PD} :

$$C_T = C_{BU} + C_{BR} + C_{PD}.$$

Since the signaling cost required for authentication and for L2 handoff are the same for all protocols ; then, they are omitted in our analysis.

3.3.3 Binding Update Signaling Cost

Depending on the type of movement and the mobility management protocol, two kinds of binding updates can be performed : *local* and *global*. For MIPv6 and FMIPv6, global binding update is performed regardless of movement every time an MN acquires a new CoA and refers to registration of CoA to HA and CNs. However, for HMIPv6, global binding update occurs when an MN moves out of its MAP domain while local binding update is performed when an MN changes its current IP address within a MAP domain. Hence, the average binding update signaling cost for IPv6-based mobility management schemes during inter-session time interval depends heavily on the computation of the number of location binding updates and is given by :

$$C_{BU} = E(N_l)C^l + E(N_d)C^g. \tag{3.7}$$

To perform signaling overhead analysis, a performance factor called session-to-mobility ratio (SMR), which represents the relative ratio of session arrival rate to the user mobility rate, is introduced. The binding update signaling cost becomes :

$$C_{BU} = \frac{1}{\lambda_s}\left(\mu_d C^g + \mu_l C^l\right) = \frac{1}{SMR\sqrt{M}}\left[C^g + (\sqrt{M}-1)C^l\right]. \tag{3.8}$$

The packet transmission cost in IP networks is proportional to the distance in hops between source and destination nodes. Furthermore, the transmission cost in a wireless link is generally larger than the transmission cost in a wired link (Xie & Akyildiz, 2002). Thus, the transmission cost of a control packet between nodes X

and Y belonging to the wired part of a network can be expressed as $C_{X,Y} = \tau d_{X,Y}$ while $C_{MN,AR} = \tau\kappa$, where τ is the unit transmission cost over wired link and κ the weighting factor for the wireless link. The global and local binding update signaling costs for MIPv6 and HMIPv6 are given by :

$$\begin{aligned} C^{g}_{MIPv6} &= C^{l}_{MIPv6} = 4C_{MN,AR} + 2PC_{AR} + C_{hc} \\ C^{l}_{HMIPv6} &= 2(2C_{MN,AR} + PC_{AR} + C_{MN,MAP}) + PC_{MAP} \end{aligned} \tag{3.9}$$

where C_{hc} is the binding update cost at the HA and at all active CNs while C_{rr} is the signaling cost due to return routability procedure. PC_{MAP} is divided into the mapping table lookup cost and the routing cost (Xie & Akyildiz, 2002).

Let consider one-way transmission cost of HoTI and CoTI messages during return routability procedure as illustrated in Fig. 3.1(a). An MN sends one HoTI message to its HA at a cost $C_{MN,HA}$. The HA processes this message at a cost PC_{HA} and forwards it to all CNs with $N_{CN}C_{HA,CN}$ as cost. Each CN processes the received HoTI message before to respond with HoT message, inducing a processing cost equal to $N_{CN}PC_{CN}$. Then, the cost for home address test is : $2[C_{MN,HA} + PC_{HA} + N_{CN}C_{HA,CN}] + N_{CN}PC_{CN}$. During the care-of address test, CoTI and CoT messages are exchanged directly between an MN and CNs. Then, the care-of address test cost is : $2N_{CN}C_{MN,CN} + N_{CN}PC_{CN}$. We can then deduce the expression of C_{rr} which is given in Table 3.2.

The link layer information (L2 trigger) is used either to predict or rapidly respond to handover events in FMIPv6. Hence, signaling cost of FMIPv6 depends on the probability that handover anticipation is correct. We assume that if an MN receives FBAck message from the PAR, then it will definitely start L3 handover to NAR without exceptions. Hence, if there is no real handover after L2 trigger, all messages exchanged from RtSolPr to FBU may be unnecessary. The local binding update

signaling cost for FMIPv6 is expressed as follows :

$$C^l_{FMIPv6} = P_s S_s + (1 - P_s)(S_f + S_r) + C_{hc} \tag{3.10}$$

where S_s denotes the signaling cost for a successfully anticipated handoff, S_f the signaling cost for control messages if no real L3 handoff occurs and S_r the signaling cost for reactive mode of FMIPv6. Their expressions are given in Table 3.2.

Tableau 3.2 Expression of partial signaling costs.

S_f	=	$3C_{MN,PAR} + 2C_{PAR,NAR} + 3PC_{AR}$
S_s	=	$4C_{MN,PAR} + 3C_{PAR,NAR} + 2C_{MN,NAR} + 5PC_{AR}$
S_r	=	$2C_{MN,PAR} + 2C_{PAR,NAR} + 2C_{MN,NAR} + 3PC_{AR}$
S^l_f	=	$3C_{MN,MAP} + 2(C_{MAP,NAR} + PC_{MAP}) + PC_{AR}$
S^l_s	=	$4C_{MN,MAP} + 3C_{MAP,NAR} + 2C_{MN,NAR} + 3PC_{MAP} + 2PC_{AR}$
S^l_h	=	$P_s[2(C_{MN,NAR} + C_{NAR,MAP}) + PC_{NAR} + PC_{MAP}] + (1 - P_s)C^l_{HMIPv6}$
C_{hc}	=	$2(C_{MN,HA} + N_{CN}C_{MN,CN}) + PC_{HA} + N_{CN}PC_{CN} + C_{rr}$
C_{rr}	=	$2(C_{MN,HA} + N_{CN}C_{HA,CN} + N_{CN}C_{MN,CN} + PC_{HA} + N_{CN}PC_{CN})$

Similar reasoning and assumption as for FMIPv6 allow computation of signaling cost for F-HMIPv6. The local binding update signaling cost of F-HMIPv6 is expressed as follows :

$$C^l_{FHMIPv6} = P_s S^l_s + (1 - P_s)S^l_f + S^l_h. \tag{3.11}$$

S^l_s and S^l_f have the same meaning as given above for FMIPv6 while S^l_h is introduced for convenient short form. Their expressions are given in Table 3.2. FMIPv6 and HMIPv6 can enhance performance of MIPv6 for movement within AN/MAP domain. However, for inter-AN/MAP movement, performance of FMIPv6 and

HMIPv6 becomes identical to that of MIPv6. If inter-MAP tunnel is not supported, the same remarks apply to F-HMIPv6.

3.3.4 Binding Refresh Cost

The binding refresh (BR) message is typically used when the cached binding is in active use but the binding's lifetime is close to expiration (Johnson *et al.* , 2004). Usually, performance analysis available in the literature did not take into account the periodic binding refresh and the effect of a binding lifetime period. However, these parameters may have significant effect on the total signaling cost. We consider it in our performance analysis and we propose the binding refresh cost. Let T_M, T_H and T_C be the binding lifetime period for the MN at MAP, HA and CNs, respectively. The average rate of sending BR message to MAP under HMIPv6 while an MN stays in a subnet is $\lfloor 1/(\mu_c T_M) \rfloor$ where $\lfloor X \rfloor$ is the integer part of a real number X. By replacing $\mu_c T_M$ with $\mu_d T_C$ and $\mu_d T_H$, respectively, we obtain average rates of sending BR message to CN and to HA. Hence, the average binding refresh costs for HMIPv6 and F-HMIPv6 can be derived as follows :

$$C_{BR}^{HMIPv6} = 2\left(\left\lfloor \frac{1}{\mu_c T_M} \right\rfloor C_{MN,MAP} + \left\lfloor \frac{1}{\mu_d T_H} \right\rfloor C_{MN,HA} + 2\left\lfloor \frac{1}{\mu_d T_C} \right\rfloor N_{CN} C_{MN,CN}\right). \tag{3.12}$$

By ignoring the binding refresh cost at MAP, we can obtain similar expression for MIPv6 and FMIPv6.

3.3.5 Packet Delivery Cost

Similarly to Koodli & Perkins (2001), we divide handoff latency into three components : link switching or L2 handoff latency (t_{L2}), IP connectivity latency (t_{IP})

and location update latency (t_U). IP connectivity latency reflects how quickly an MN can send IP packets after L2 handoff while location update latency is the latency of forwarding IP packets to MN's new IP address. On the other hand, the time from the starting point of L2 handoff to when an MN first receives IP packets for the first time after link switching refers to packet reception latency (t_P) or handoff latency. Moreover, we define the following delay components : movement detection delay (t_{MD}), addresses configuration and DAD procedure delay (t_{AC}), binding update latency (t_{BU}) and delay from completion of binding update and reception of first packet at the new IP address (t_{NR}).

Fig. 3.4 illustrates the timing diagram associated to MIPv6 and shows that there is a delay before an MN begins to receive packets directly through the NAR. The packet delivery cost incurs during ongoing session and is composed of packet

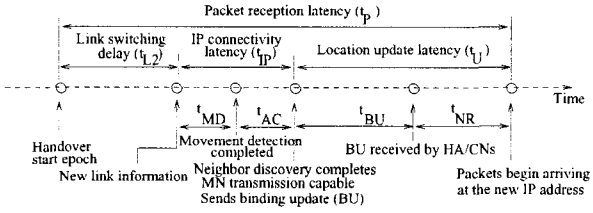

Figure 3.4 Handoff delay timeline of MIPv6.

transmission and processing costs. The packet delivery cost could be defined as the linear combination of packet tunneling cost (C_{tun}) and packet loss cost (C_{loss}). Let α and β be weighting factors (where $\alpha + \beta = 1$), which emphasize tunneling effect and dropping effect ; then, the packet delivery cost is computed as follows :

$$C_{PD} = \alpha C_{tun} + \beta C_{loss}. \tag{3.13}$$

Let s_c and s_d be the average size of control packets and data packets, respectively and $\eta = s_d/s_c$. The cost of transferring data packet is η greater than the cost of

transferring control packet. Let λ_p be the packet arrival rate in unit of packet per time. There is no forwarding with MIPv6 during handover (i.e., $C_{tun}^{MIPv6} = 0$) ; then, only packet loss cost occurs and is evaluated as follows :

$$C_{loss}^{MIPv6} = \lambda_p C_{cm}^{f,1}(t_{L2} + t_{IP} + t_U) \tag{3.14}$$

where $C_{cm}^{f,1} = \eta(C_{CN,PAR} + C_{PAR,MN})$ is the cost of transferring data packets from CN to MN via PAR when the handoff fails, $t_U = t_{BU} + t_{NR}$, $t_{BU} = t_{HA} + t_{RR} + t_{CN}$, t_{HA} is the delay for performing BU process to the HA, t_{RR} is the delay for return routability procedure and t_{CN} is the delay of BU procedure to all active CNs.

In HMIPv6, all packets directed to MN will be received by MAP and after being tunneled to MN's current address (LCoA) by using mapping table. Then, the lookup time of mapping table has an effect on MAP's processing cost. Similarly to MIPv6, there is no forwarding with HMIPv6 during handover (i.e., $C_{tun}^{HMIPv6} = 0$). Hence, packet delivery cost for intra-AN/MAP roaming can be computed through (3.13), where packet loss cost, C_{loss}^{HMIPv6}, is given by :

$$C_{loss}^{HMIPv6} = \lambda_p C_{cm}^{f,2}(t_{L2} + t_{IP} + t_U^L) \tag{3.15}$$

where t_U^L is the location update latency for intra-AN/MAP roaming : $t_U^L = t_{BU}^L + t_{NR}^L$ with t_{BU}^L the local binding update latency at MAP while t_{NR}^L is equivalent to t_{NR} for local roaming and the transferring data packets cost between CN and MN when the handoff fails is $C_{cm}^{f,2} = \eta(C_{CN,MAP} + C_{MAP,PAR} + C_{PAR,MN} + PC_{MAP})$.

To avoid packet loss, FMIPv6 enables PAR to forward packets to NAR by using a bi-directional tunnel established between them and by buffering all forwarded packets. The timing diagram of predictive mode of FMIPv6 is shown in Fig. 3.5

and the packet tunneling cost is given by :

$$C_{tun}^{FMIPv6,p} = \lambda_p C_{cn}^{s,1}(t_{L2} + t_{IP}^P + t_U) \tag{3.16}$$

where $C_{cn}^{s,1} = \eta(C_{CN,PAR} + C_{PAR,NAR} + C_{NAR,MN})$ is the cost of transferring data packets from CN to MN by transiting to PAR and forwarding to NAR via the established tunnel, and t_{IP}^P is the IP connectivity latency for predictive mode of fast handover scheme, $t_{IP}^P \leq t_{IP}$.

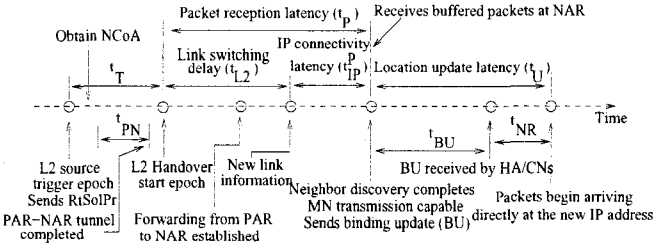

Figure 3.5 Handoff delay timeline of FMIPv6.

The packet loss due to L2 handoff delay is inevitable without an efficient buffering mechanism. Moreover, packet loss in FMIPv6 may be due to wrong temporal and spatial predictions. Let t_{PN}, be the time required to establish a tunnel between PAR and NAR. Usually, t_T is greater than t_{PN} ; then, packets received during handover procedure are forwarded by PAR to NAR by using the already established tunnel. But, if MN moves very fast, t_T may be less than t_{PN}. Then, packets arriving to PAR during the time period $t_{PN} - t_T$ may be lost, because the tunnel is not yet established. In other words, for the anticipated signaling to succeed, the following time constraint must be observed : $t_{PN} \leq t_T$. Hence, packet loss cost for predictive mode of FMIPv6 can be expressed as follows :

$$C_{loss}^{FMIPv6,p} = \lambda_p C_{cn}^{f,1} \max(t_{PN} - t_T, 0). \tag{3.17}$$

Due to wrong spatial prediction of NAR or if FBAck message was not received on the previous link, the forwarded packets by PAR may be lost. In this case, the reactive mode of FMIPv6 is used. Let t_{IP}^{R}, the IP connectivity latency of reactive mode. Since the packets forwarding process is not supported in the reactive mode; then, packet tunneling cost is equal to zero while packet loss cost for reactive mode of FMIPv6 can be expressed as follows :

$$C_{loss}^{FMIPv6,r} = \lambda_p C_{cm}^{f,1}(t_{L2} + t_{IP}^{R} + t_U). \tag{3.18}$$

Hence, the average packet delivery cost of FMIPv6 in terms of prediction accuracy is given by :

$$C_{PD}^{FMIPv6,a} = P_s C_{PD}^{FMIPv6,p} + (1 - P_s) C_{PD}^{FMIPv6,r}. \tag{3.19}$$

With similar reasoning to FMIPv6, evaluation of packet delivery cost for intra-AN/MAP roaming for F-HMIPv6 is obtained by replacing t_U, t_{PN}, t_{IP}^{R}, $C_{cm}^{s,1}$ and $C_{cm}^{f,1}$, respectively by t_U^L, t_{ML}, t_{IP}, $C_{cm}^{s,2}$ and $C_{cm}^{f,2}$. Where $C_{cm}^{s,2}$ is the cost of transferring data packets from CN to MN by transiting through the MAP and NAR given by $C_{cm}^{s,2} = \eta(C_{CN,MAP} + C_{MAP,NAR} + C_{NAR,MN} + PC_{MAP})$ and t_{ML} is the time required to establish a tunnel between MAP and NAR. For inter-AN/MAP roaming, the packet delivery cost of HMIPv6, FMIPv6 and F-HMIPv6 becomes the same as for MIPv6.

3.3.6 Required Buffer Space

In FMIPv6, the NAR buffers packets tunneled from the PAR and forwards them to MN when the latter announces its presence on the new link. Hence, the required buffer space during MN's subnet movement increases in proportion of the

packet arrival rate and according to the number of MNs performing handover. The buffer space required for FMIPv6 during intra-AN/MAP handover is proportional to handoff latency and is computed as follows :

$$BS^{l}_{FMIPv6} = \lambda_p[P_s(t_{L2} + t^{P}_{IP} + t_U) + (1 - P_s)t_{NR}]. \tag{3.20}$$

Similarly, buffer space required for F-HMIPv6 is obtained by replacing t_U and t_{NR} by t^{L}_{U} and t^{L}_{NR} in (3.20), respectively. Since MIPv6 and HMIPv6 do not use handover anticipation techniques; then, by setting $P_s = 0$ in (3.20), we obtain a required buffer space for MIPv6 and HMIPv6.

3.3.7 Handoff Latency and Packet Loss

We define the following parameters to compute handoff latency and packet loss : t_{L2} the L2 handoff latency, t_{RD} the round-trip time for router discovery procedure, t_{DAD} the time for DAD process execution, t_{RR} the delay for an MN to perform return routability procedure and $t_{X,Y}$ one-way transmission delay of a message of size s between nodes X and Y. Since the average delay needed for an MN authentication is the same for all protocols; then, it is omitted. If one of the endpoints is an MN, $t_{X,Y}$ is computed as follows :

$$t_{X,Y}(s) = \frac{1-q}{1+q}\left(\frac{s}{B_{wl}} + L_{wl}\right) + (d_{X,Y} - 1)\left(\frac{s}{B_w} + L_w + \varpi_q\right) \tag{3.21}$$

where q is the probability of wireless link failure, ϖ_q the average queueing delay at each router in the Internet (McNair *et al.* , 2001), B_{wl} (resp. B_w) the bandwidth of wireless (resp. wired) link and L_{wl} (resp. L_w) wireless (resp. wired) link delay. The handoff latency associated to MIPv6 is given by :

$$D_{MIPv6} = t_{L2} + t_{RD} + t_{DAD} + t_{RR} + 2(t_{MN,HA} + t_{MN,CN}). \tag{3.22}$$

The handoff latency for intra-AN/MAP or localized movement of HMIPv6 is obtained by replacing HA by MAP and by ignoring t_{RR} and $t_{MN,CN}$ in (3.22). Let Δ_{ns} be the time elapsed from the reception of FBAck on previous link to the beginning of L2 handoff when there is no good synchronization between L2 and L3 handoff mechanisms. Moreover, let Δ_{lr} be the time between last packet reception through previous link and L2 handoff beginning when FBAck is received on new link. Note that, Δ_{lr} and Δ_{ns} may be equal to zero and we use this assumption in performance analysis. For fast handoff schemes, the handoff latency depends on information availability, and on which link fast handoff messages are exchanged. Hence, if information about NAR and impending handoff are available, and FBAck message is received through the previous link, handoff latency for localized or micro-mobility without an efficient buffers management for FMIPv6 and F-HMIPv6 is expressed as follows :

$$O^l_{FMIPv6} = O^l_{FHMIPv6} = \Delta_{ns} + t_{L2} + 2t_{MN,NAR}. \tag{3.23}$$

If FBAck message is not received through previous link, F-HMIPv6 turns to HMIPv6 while for FMIPv6 its reactive mode is used. Then, handoff latency without efficient buffer management for FMIPv6 is expressed as follows :

$$N^l_{FMIPv6} = \Delta_{lr} + t_{L2} + 2t_{MN,NAR} + 3t_{NAR,PAR}. \tag{3.24}$$

The average handoff latency for FMIPv6 is expressed as follows :

$$D^l_{FMIPv6} = P_s O^l_{FMIPv6} + (1 - P_s) N^l_{FMIPv6}. \tag{3.25}$$

Similarly, we can obtain the average handoff latency for F-HMIPv6. The predictive mode of FMIPv6 cannot perform anticipated IP-handoff for inter-AN (Gwon *et al.*

, 2004) ; then handoff latency of FMIPv6 becomes same as for MIPv6. The same remark applies to HMIPv6 and F-HMIPv6.

With MIPv6 and HMIPv6, packet loss occurs during handoff latency or service disruption latency. In fact, the number of packet loss is proportional to handoff latency. This is also the case for FMIPv6 and F-HMIPv6 if there is no efficient buffer management (BM). In fact, for fast handoff schemes there is no packet loss in theory, unless buffer overflow happens. Hence, the number of packet lost for each handoff management scheme is computed as follows :

$$P_{loss}^{scheme,l} = \begin{cases} \max(BS_{scheme}^{l} - B, 0) & \text{for efficient BM} \\ \lambda_p D_{scheme}^{l} & \text{otherwise} \end{cases} \tag{3.26}$$

where B is the buffer size of an AR and BS_{scheme}^{l} is the buffer space required at an access router for a given scheme (i.e., MIPv6, HMIPv6, FMIPv6 or F-HMIPv6).

3.4 Performance Evaluation

Parameters and default values used in performance evaluation are given in Table 3.3, except when wireless link delay and packet arrival rate are considered as variable parameters. The network topology considered for analysis is illustrated in Fig. 3.6, where ER means edge router. For protocols which do not involve hierarchical mobility management, the MAPs act as a normal intermediate (edge) router. We assume that distance (i.e., the number of hops) between different domains are equals, i.e., $c = d = e = f = 10$ and we set $a = 1$, $b = 2$. The time-to-live (TTL) field in IP packet headers may be used by an MN to get the number of hops packets travel. Then, this distance varies within a certain range (Xie & Akyildiz, 2002). All links are supposed to be full-duplex in terms of capacity and delay. Other para-

Tableau 3.3 System parameters.

Parameters	Symbols	Values
DAD delay	t_{DAD}	500 ms
Router discovery delay	t_{RD}	100 ms
L2 handoff delay	t_{L2}	50 ms
Prediction probability	P_s	0.90
Wireless link failure probability	q	0.50
Wired link bandwidth	B_w	100 Mbps
Wireless link bandwidth	B_{wl}	11 Mbps
Wired link delay	L_w	2 ms
Wireless link delay	L_{wl}	10 ms
Number of ARs by AN/MAP	M	2
Control packet size	s_c	96 bytes
Data packet size	s_d	200 bytes
Packet arrival rate	λ_p	10 packets/s
MN average speed	v	5.6 Km/h
Subnet radius	R	500 m

meters used for cost computation are defined as follows : $\tau = 1$, $\kappa = 10$, $\alpha = 0.2$, $\beta = 0.8$, $\sigma = 2$, $PC_{AR} = 8$, $PC_{HA} = 24$, $PC_{CN} = 4$ and $PC_{MAP} = 12$. Most parameters used in this analysis are set to typical values found in Xie & Akyildiz (2002); Pack & Choi (2003) and Lai & Chiu (2005).

Fig. 3.7 illustrates the binding update signaling cost during handoff as a function of SMR for intra-AN/MAP roaming. When SMR is small, the mobility rate is larger than session arrival rate ; then, an MN changes subnet frequently due to its mobility, inducing several handoffs and the signaling overhead increases. However, when the session arrival rate is larger than mobility rate (i.e., SMR is greater than 1), binding update is less often performed and signaling overhead decreases because the frequency of subnet changes decreases. FMIPv6 and F-HMIPv6 do not effectively reduce signaling overhead comparatively to MIPv6 and HMIPv6, respectively due to messages introduced for handoff anticipation. However, signaling overhead of

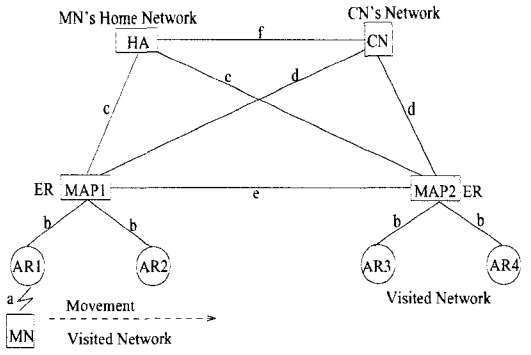

Figure 3.6 Network topology used for analysis.

fast handoff schemes is traded off by lower handoff latency and packet loss as we will see later.

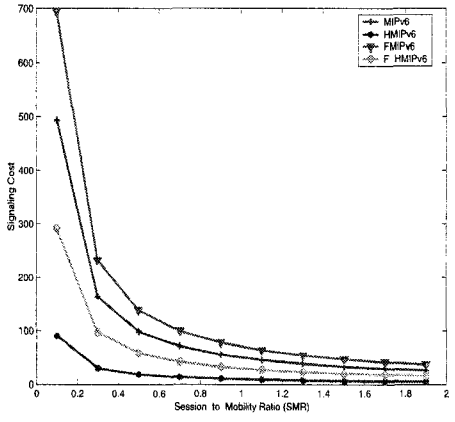

Figure 3.7 Impact of session-to-mobility ratio on binding update.

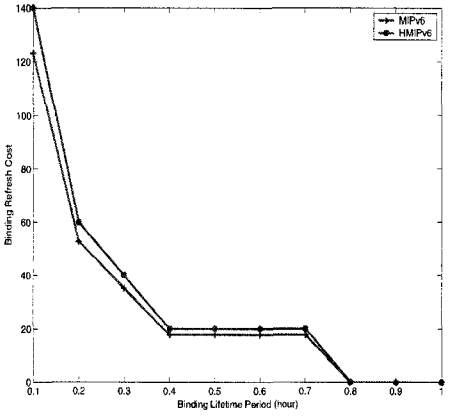

Figure 3.8 Impact of binding lifetime period on binding refresh cost.

Fig. 3.8 represents the effect of binding lifetime period on the binding refresh cost and shows that the binding refresh cost decreases as binding lifetime period increases. We assume that the binding lifetime periods T_M, T_H and T_C are equals. We can see that the binding lifetime period has significant impact on the average binding refresh cost. Small value of binding lifetime period leads to larger binding refresh cost; in other words, significant signaling load throughout the network. On

the other hand, larger value of the binding lifetime period leads to larger binding cache entry at mobility agents. This may result in higher memory consumption and higher binding cache lookup time.

The result shows that the binding refresh cost remains constant when the binding lifetime period is between 0.4 and 0.7 hour and as well as when it is greater than 0.8 hour. For the former case, the result indicates that during $[0.4, 0.7]$ time period there is the same number of binding refresh messages. This is due to the fact that an MN moves to an adjacent subnet before the new binding refresh message occurs. While for the latter case, the average subnet residence time of an MN is shorter than the binding lifetime period (i.e., $T_M \geq 0.8$ hour). Hence, no binding refresh message occurs and the binding refresh cost is equal to zero. On the other hand, due to binding cache and lookup table maintained at the MAP, there is an extra cost for binding refresh process at the MAP for HMIPv6. Thus, binding refresh cost of HMIPv6 is slightly greater than for MIPv6.

The packet delivery cost is depicted in Fig. 3.9 as a function of packet arrival rate (λ_p). We observe that, packet delivery cost increases proportionally with λ_p for all schemes. Fast handoff schemes (i.e., F-HMIPv6 and FMIPv6) outperform MIPv6 and HMIPv6, and they are more efficient when λ_p increases. This means that FMIPv6 and F-HMIPv6 are better suited for real-time applications where periodic packets are sent at high rates. The packet delivery cost depends on handoff latency, while packet loss is proportional to handoff latency. Then, a similar analysis may be performed for packet loss when comparing to packet arrival rate as in Fig. 3.9. Hence, packet loss will be lesser for fast handoff schemes than for MIPv6/HMIPv6.

For varying prediction probability, P_s, Fig. 3.10 shows the behavior of packet delivery cost. The packet delivery cost decreases when the accuracy of P_s increases for

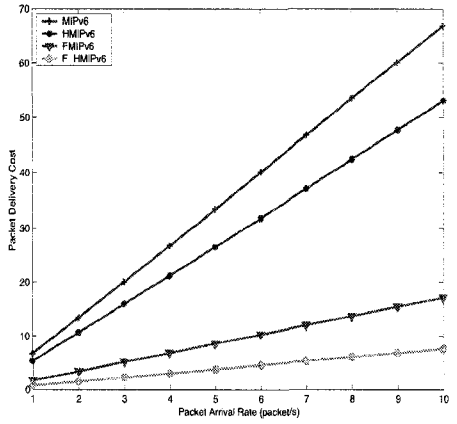

Figure 3.9 Packet delivery cost as a function of packet arrival rate.

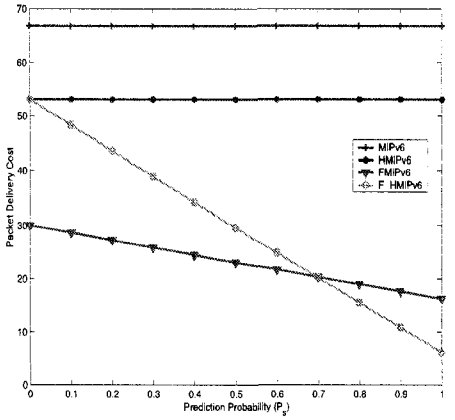

Figure 3.10 Packet delivery cost as a function of prediction probability.

fast handoff schemes. Due to additional packet processing at MAP for F-HMIPv6, there is an extra cost for packet delivery with inaccuracy prediction. In fact, in this case, F-HMIPv6 turns to HMIPv6, as we can see when $P_s = 0$. HMIPv6 and MIPv6 are not affected by the prediction probability. For high values of P_s, F-HMIPv6 performs better than FMIPv6. Since there is a relation between handoff latency and packet delivery cost, a similar behavior will be observed when comparing handoff latency with prediction probability. Hence, an effective prediction mechanism is required to allow better performance for F-HMIPv6.

To alleviate packet losses, fast handover schemes should support packet buffering and forwarding during handoff execution. Since fast handover schemes start packet buffering and forwarding earlier ; then, they require more buffer space than MIPv6 and HMIPv6 as we can see in Fig. 3.11. On the other hand, buffering time may affect real-time applications, for example if some packets are stored in a buffer for a longer period of time than acceptable end-to-end delay, they may become useless. Hence, it is crucial to manage buffers efficiently in order to minimize overhead and to provide better QoS to delay sensitive applications.

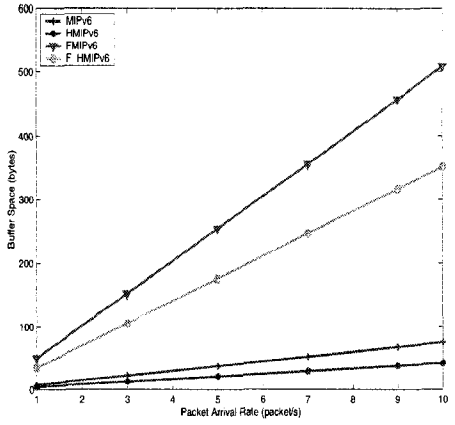

Figure 3.11 Required buffer space as a function of packet arrival rate.

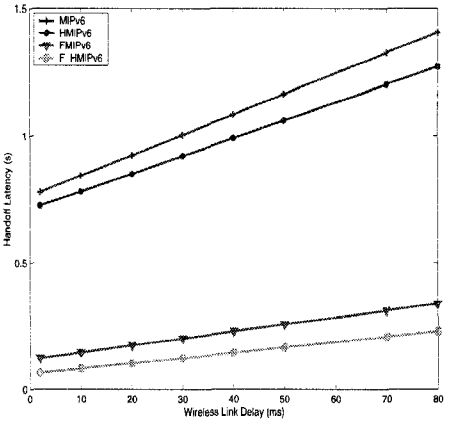

Figure 3.12 Impact of wireless link delay on handoff latency.

In Fig. 3.12, we can see that the handover latency increases proportionally with the wireless link delay. We observe that MIPv6 and HMIPv6 have worst results among all protocols followed by FMIPv6 while F-HMIPv6 performs better than all other schemes. For MIPv6 and HMIPv6, the DAD process counts for a large portion of handoff delay. Therefore, it is important to decrease the DAD delay in order to decrease handoff latency. The optimistic DAD (oDAD) (Moore, 2006) has recently been proposed to allow minimization of address configuration delay by eliminating the DAD completion time.

3.5 Conclusion

Mobility management is a key issue in next-generation or 4G wireless networks (NGWN/4G). Several IPv6-based mobility schemes have been proposed in the literature and by the IETF. However, they are not able to guarantee seamless roaming and services continuity for critical applications like real-time applications. Moreover, performance evaluation of these schemes is usually based on simulation approaches.

This paper proposes a comprehensive analytical model for IPv6-based mobility protocols (i.e., MIPv6, HMIPv6, FMIPv6 and F-HMIPv6) in order to provide depth analysis of the overall performance of these protocols. Several performance metrics such as signaling overhead cost, packet delivery cost, handoff latency and packet loss are analyzed according to user mobility and traffic models. Our goal was not to decide which scheme is always better, but to study the effect of various parameters related to mobility and traffic on the performance of these schemes in order to facilitate decision-making for wireless network design.

The numerical results show the potential pros and cons of most promising IPv6-based mobility schemes proposed by the IETF. They reveal that F-HMIPv6 enables improvement in terms of handoff latency and packet loss rather than other protocols (i.e., MIPv6, HMIPv6 and FMIPv6). However, this performance is off-set by its signaling traffic overhead and the buffer space required when compared to HMIPv6. Moreover, it is very difficult to forecast which IPv6-based mobility protocol will dominate in NGWN/4G. In fact, selection of a mobility management scheme is not based solely on performance criteria, but on cost and respective profits as well. Thus, until an ideal mobility management protocol is designed and deployed, mobile users still require a practical solution. This could be achieved by a certain tradeoff of the above requirements.

CHAPITRE 4

AN ARCHITECTURE FOR SEAMLESS MOBILITY SUPPORT IN IP-BASED NEXT-GENERATION WIRELESS NETWORKS

Christian Makaya and Samuel Pierre
Mobile Computing and Networking Research Laboratory (LARIM)
Department of Computer Engineering, École Polytechnique de Montréal
P.O. Box 6079, Station Centre-ville, Montréal, Québec, H3C 3A7, Canada
Email : {christian.makaya, samuel.pierre}@polymtl.ca

Abstract

Recent technological innovations allow mobile devices to be equipped with multiple wireless interfaces. Moreover, the trend in the fourth generation or next-generation wireless networks (4G/NGWN) is the coexistence of diverse but complementary architectures and wireless access technologies. In this context, an appropriate mobility management scheme as well as the integration and interworking of existing wireless systems are crucial. Several proposals are available in the literature to solve these issues. However, these proposals cannot guarantee seamless roaming and services continuity. This paper proposes a novel architecture, called *Integrated InterSystem Architecture* (IISA), based on a 3GPP/3GPP2 proposal, which enables the integration and interworking of current wireless systems and investigates mobility management issues. An efficient handoff protocol based on localized mobility management, access networks discovery and fast handoff concepts, called *Handoff Protocol for Integrated Networks* (HPIN) is proposed. It alleviates services disruption during handoff in IPv6-based heterogeneous wireless environments. Per-

formance evaluation based on numerical results shows that HPIN performs better in terms of signaling cost, handoff latency, handoff blocking probability and packet loss compared to existing schemes.

Keywords : Mobility management, quality of service, IP-based wireless networks, vertical handoff, interworking architecture, seamless roaming, service continuity.

4.1 Introduction

Next-generation or 4G wireless networks (NGWN/4G) are expected to exhibit heterogeneity in terms of wireless access technologies, services, application requirements, high usability and improved capacity. With NGWN/4G, users will intensify demands for seamless roaming across different wireless networks, support of various services (e.g., multimedia applications) and quality of service (QoS) guarantees. The strengths of 3G cellular networks, such as UMTS and CDMA2000, consist of their global coverage while their weaknesses lie in bandwidth capacity and operation costs. On the other hand, WLAN technology, such as IEEE 802.11, offers higher bandwidth with low operation costs, although it covers a relatively short range. Moreover, technological advances in evolution of portable devices have made possible the support of different radio access technologies (RATs).

This has raised much interest in integration and interworking of 3G wireless networks with WLAN due to the potential benefits of their complementarity. Evolution through this integration is one of the paths to NGWN design, rather than investing efforts into developing new radio interfaces and technologies (Hui & Yeung, 2003). Integrated networks will provide benefits of both technologies to end-users as well as to services providers. The integration of wireless networks will not be limited only to WLAN and 3G cellular networks but will be extended to other networks

also such as satellite networks, WiMAX, mobile ad hoc networks, wireless sensor networks, etc.

Conceptually, a typical NGWN architecture can be viewed as many overlapping wireless access domains, as shown in Fig. 4.1 and is so-called wireless overlay networks (Stemm & Katz, 1998). The main goal of NGWN is to allow subscribers to profit services anytime and anywhere, known as *always best connected* (Gustafsson & Jonsson, 2003). The heterogeneity in terms of RATs and network protocols in NGWN asks for common interconnection element. Since the Internet Protocol (IP) technology enables the support of applications in a cost-effective and scalable way, it is expected to become the core backbone network of NGWN (Akyildiz *et al.* , 2005). Hence, current trends in communication networks evolution are directed towards the *all-IP* principle in order to hide heterogeneities and achieve convergence of various networks.

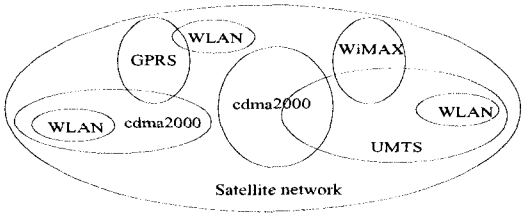

Figure 4.1 Overview of 4G/NGWN architecture.

Two major architectures (loose and tight coupling) for 3G/WLAN interworking based on existing 3G network architecture components have been proposed in 3GPP (2004). All scenarios presented in 3GPP (2004); 3GPP2 (2006) are not yet fulfilled and those interworking architectures have pros and cons. The integration and interworking of heterogeneous wireless networks are widely documented in the literature and various models have been proposed. Both 3G wireless networks initiatives, 3GPP and 3GPP2, have proposed a 3G/WLAN interworking architecture adapted to their respective systems. An evident way to achieve roaming among

various networks is by using bilateral Service Level Agreements (SLAs). However, due to several reasons this approach is not feasible. In fact, the increasing number of wireless networks and service providers make it impractical for network operators to have direct SLAs with all of the other operators. Moreover, network operators are reticent to make their databases available to other operators.

Mobility management, with provision of seamless handoff and QoS guarantees, consist of one of the key issues in order to support global roaming of mobile nodes (MNs) between various wireless systems in an efficient way. In NGWN, mobility is also a logical concept rather than only a physical one. It is thus crucial to provide seamless roaming and QoS guarantee support based on intelligent and efficient mobility management schemes. To enable services continuity and QoS provision, seamless handoff (i.e., minimal services disruption during handoff) is of great importance. Seamless handoff means lower packet loss, minimal handoff latency, lower signaling traffic overhead and limited handoff failure. The handoff latency refers to the time interval during which an MN cannot send or receive any data traffic during handoffs. It is composed of L2 (link layer) and L3 (IP layer) handoff latency. The overall handoff latency may be sufficiently long to cause packets loss, which is unacceptable for real-time applications.

The QoS guarantee represents one of the major challenging issues due to the heterogeneity of network architectures, network capacities, different high layer protocols and various radio access technologies (RATs). An exact mapping between all 3G wireless network QoS parameters and WLAN QoS parameters is highly difficult to perform and remains an open issue, since these networks are totally different. The handoff process in NGWN can be subdivided into three phases : network discovery, handoff decision and handoff execution. The simplest way for an MN with multiple air-interfaces to discover reachable wireless networks is to keep all air-interfaces on at all times. However, keeping an air-interface active continuously consumes

battery power even while the mobile device is not sending/receiving packets. It is thus critical to avoid keeping idle air-interfaces perpetually on. Moreover, an MN must observe if the new network is consistently better than the current one before performing handoff, to avoid the ping-pong effects.

In homogeneous wireless networks, handoff decisions are typically driven by metrics strictly related to received signal strength (RSS) quality and resources availability. However, in NGWN, RSSs from different networks do not share the same meaning since each network is composed of its specific features; then, it cannot be compared directly. Hence, handoff decisions based on signal strength as the sole criterion may be inefficient or impractical in NGWN. More complex metrics combining several parameters such as monetary costs, bandwidth, power consumption, network conditions and user preferences must be defined (McNair & Zhu, 2004).

This paper proposes a novel architecture, called *Integrated InterSystem Architecture* (IISA), based on 3GPP/3GPP2-WLAN interworking models, in order to integrate existing wireless systems such as 3GPP/3GPP2, WLAN and WiMAX, and hide their heterogeneities. Furthermore, we propose a mobility management scheme, called *Handoff Protocol for Integrated Networks* (HPIN), that provides QoS guarantee for real-time applications in heterogeneous IPv6-based wireless environments. HPIN is a one-suite protocol that performs access networks discovery, and uses fast handoff and localized mobility management concepts. HPIN allows the selection of the best available network at any given time and it is designed for both heterogeneous and homogeneous wireless networks. In other words, the main contributions of this paper are as follows :

1)- the design of an interworking architecture that permits integration of any type of wireless networks rather than only 3G cellular systems with WLAN or heterogeneous 3G cellular systems;

2)- the design of an efficient handoff management scheme which enables the support of seamless handoff and services continuity for mobile users moving across various networks;

3)- the proposal of a new approach to speed up context transfers and binding updates for mobile users;

4)- the proposal of an analytical model to analyze the performance of the proposed mechanisms and architecture.

The remainder of this paper is organized as follows. The following section offers an overview of the basic concepts and related work pertaining to interworking and mobility management in heterogeneous wireless networks. Then, the proposed architecture (IISA) and the handoff management scheme (HPIN) are described respectively in Sections 4.3 and 4.4. The analytical model is developed in order to assess their efficiency in Section 4.5. Results from the performance evaluation are analyzed in Section 4.6 before concluding remarks drawn in the last section.

4.2 Background and Related Work

Mobility management enables a system to locate roaming terminals in order to deliver data packets (i.e., *location management*) and maintain connections with them when moving into a new subnet (i.e., *handoff management*). Handoff management is a major component of mobility management since an MN can trigger several handoffs during a session as it will be the case in NGWN. Handoffs in IP-based NGWN involve changes of access points or base stations (AP/BSs) at the link layer and possibly routing changes at the IP layer. With the coexistence of various wireless access technologies, two kinds of handoffs are possible in NGWN : *horizontal* and *vertical handoffs*. Horizontal or intrasystem handoffs occur when an MN is moving between AP/BSs of the same network technology. When AP/BSs

belong to different networks (e.g., IEEE 802.11 and UMTS), such a movement is called a vertical handoff.

Two types of vertical handoffs can occur depending on the type of overlapping. In fact, roaming may happen between fully overlapping networks from low-tier (e.g., WLAN) to high-tier (e.g., 3G wireless network) and vice versa, or between partially overlapping networks. In case of roaming under fully overlapping networks, vertical handoffs are usually asymmetric and can focus on improving either the transmission rate or session connectivity (Stemm & Katz, 1998). The characteristics of NGWN make the implementation of vertical handoffs more challenging than horizontal handoffs. In fact, maintaining uninterrupted sessions while the physical interface is changing constitutes a complex task. Several IPv6-based handoff protocols proposed in the literature in order to manage horizontal and vertical handoffs may appear appropriate. However, they have advantages and drawbacks and have been proposed separately. Much works still required for further improvements in NGWN/4G.

4.2.1 IPv6-based Mobility Schemes

Mobile IPv6 (MIPv6) was proposed by the Internet Engineering Task Force (IETF) for mobility management at the IP layer and allows MNs to remain reachable in spite of their movements within wireless IP environments (Johnson *et al.* , 2004). MNs are always identified by their home address, regardless of their current network point of attachment. While away from its home network, an MN is associated with a care-of address (CoA), which provides information about its current location. After acquiring a CoA, an MN sends a binding update (BU) message to the home agent (HA), to indicate its new address and also to all active correspondent nodes (CNs) to allow route optimization. However, MIPv6 has some well-known drawbacks such as signaling traffic overhead, high packet loss rate and

handoff latency, thereby causing user-perceptible deteriorations of real-time traffic (Pérez-Costa *et al.* , 2003; Gwon *et al.* , 2004).

These weaknesses led to the investigation of other solutions to enhance MIPv6. Two main MIPv6 extensions proposed by the IETF are Fast Handovers for MIPv6 (FMIPv6) (Koodli, 2005) and Hierarchical MIPv6 (HMIPv6) (Soliman *et al.* , 2005). These protocols tackle micro-mobility while MIPv6 is used for macro-mobility. HMIPv6 handles handoff locally through a special node called Mobility Anchor Point (MAP). The MAP, acting as a local HA in the network visited by the MN, limits the amount of MIPv6 signaling outside its domain and reduces delays associated with the location updates. However, HMIPv6 cannot meet the requirements for delay-sensitive traffic, such as voice over IP (VoIP), due to packets loss and handoff latency. FMIPv6 has been proposed in order to minimize services disruption during handoffs pertaining to MIPv6 operations such as movement detection, binding update and addresses configuration. The link layer information (L2 trigger) is used either to predict or to respond rapidly to handoff events.

Although FMIPv6 paves the way to improve MIPv6 performance in terms of handoff latency, it is still hindered by several problems such as QoS support and scalability. In fact, FMIPv6 does not effectively reduce signaling overhead nor packet loss, which leads to unacceptable services disruption. In FMIPv6, the new access router (NAR) consumes storage space to buffer forwarded packets by previous access router (PAR) before delivering these packets to the MN. These forwarded packets lack QoS guarantee before the new QoS path is set up. Combining HMIPv6 and FMIPv6 motivates the design of Fast Handover for HMIPv6 (F-HMIPv6) (Jung *et al.* , 2005a) to increase network bandwidth usage efficiency. However, F-HMIPv6 may inherit drawbacks of both FMIPv6 and HMIPv6, such as synchronization issues and signaling overhead (Pérez-Costa *et al.* , 2003; Gwon *et al.* , 2004). With those IPv6-based mobility protocols, seamless mobility cannot be guaranteed.

To achieve seamless mobility across various access technologies and networks, an MN needs information about the wireless network to which it could attach. Also, it is necessary to transfer information (context transfer) related to an MN from the current access router to the next one. To enable these procedures, the Candidate Access Router Discovery protocol (CARD) (Leibsch *et al.* , 2005) and the Context Transfer Protocol (CXTP) (Loughney *et al.* , 2005) have been proposed. They avoid using limited wireless resources and provide fast mobility and secure transfers. Their key objectives consist of reducing latency, packet losses, and avoiding the re-initiation of signaling to/from an MN from the beginning during an handoff. However, context transfer is not always possible. For example, when an MN moves across different administrative domains, the new network may require the MN to re-authenticate and perform signaling from beginning rather than accepting the transferred context. With the CARD protocol, acquiring L3 information of neighbor ARs is based on L2 ID detection, which is possible only when the associated air-interface is on. Also, MNs must periodically monitor the RSS from neighbor AP/ARs and construct neighbor network information table. Moreover, entities exchanging contexts must authenticate each other, which could turn into a tedious procedure in 4G/NGWN.

4.2.2 3G/WLAN Interworking Models

Six 3G/WLAN interworking scenarios and their requirements have been defined in 3GPP (2003) and 3GPP2 (2004) in order to provide a proper background for interworking architecture design. With the particular characteristics of WLAN and 3G wireless networks, two scenarios present significant technical challenges : services continuity and seamless roaming provision. In order to handle these scenarios, two interworking architectures have been proposed by 3GPP, called *loose* and *tight coupling* (3GPP, 2004). With the tight coupling approach, WLAN appears as one

of the 3G radio access networks (RANs) to the 3G wireless core network. Although, the tight coupling allows easy control of QoS for time-sensitive traffic, it includes several drawbacks such as high costs and complexity levels. Moreover, with tight coupling, traffic from WLAN flows into 3G wireless core network and it creates capacity problems. In fact, 3G wireless core network nodes cannot accommodate the bulk of the data traffic from WLAN.

On the other hand, with the loose coupling, different networks are deployed independently and data paths are completely separated between WLAN and 3G wireless networks. Hence, loose coupling enables several advantages such as independent traffic engineering, low costs and low complexity levels. However, loose coupling may not guarantee services continuity to other access networks during handoff as it suffers from long handoff latency and packet loss. The choice of an optimal interworking architecture is determined by a certain number of factors. For example, if the wireless network is composed of a large number of WLAN and 3G networks operators, the loosely coupled architecture would be the best choice. On the other hand, if the WLAN network is exclusively owned by 3G wireless operator, the tightly coupled architecture might become a more relevant option. However, loose coupling is the most advocate interworking scheme (Buddhikot *et al.* , 2003).

4.2.3 Handoff Management Schemes

In Buddhikot *et al.* (2003), an integrated architecture and a radio interface selection schemes are proposed based on signal strength and radio interface priorities. As aforementioned, these parameters are not appropriate for handoff decisions in NGWN. Moreover, an MN must passively evaluate handoff conditions, even when the application in the current network is running well. This introduces unnecessary power consumption and network resources usage. HOPOVER (HandOff Protocol

for OVERlay networks), a mobile IP-based approach was proposed in Du *et al.* (2002) and handles both vertical and horizontal handoffs. Although, HOPOVER enables low signaling overhead, it requires APs to maintain an excessive quantity of information about MNs. An architecture for next generation all-IP-based wireless systems was proposed in Akyildiz *et al.* (2005). Two new entities, the network interworking agent (NIA) and the interworking gateway (IG), are introduced in order to allow the integration of several wireless networks while supporting MN roaming. However, this proposed architecture provides no appropriate handoff decision mechanism to take heterogeneity into account. The handoff decision is based on RSS criterion, which as mentioned above is not appropriate for NGWN.

In Wang & Akyildiz (2001), a mobility management scheme and an architecture are proposed to support roaming across 3G heterogeneous wireless networks but not for IP-based wireless networks or *authentic* NGWN. This proposed architecture is based on a boundary location register (BLR) and a border interworking unit (BIU), which are placed at the border of two neighboring systems. This approach is not scalable in the sense that one BLR/BIU is needed for each pair of adjacent networks. Furthermore, connecting directly BIUs to Visitor Location Register (VLR) of each subsystem creates several drawbacks, such as the increase of cabling costs and signaling traffic due to paging procedure execution through the Home Location Register (HLR) of both involved networks.

The BLR/BIU model cannot meet all of the main requirements (economics, scalability, transparency to heterogeneous access technologies, seamless mobility support and security) of any novel architecture. An architecture and mobility management scheme that improve performance of BLR/BIU model was proposed in Beaubrun *et al.* (2005). Although, an HLR and a VLR may be seen as home agent (HA) and foreign agent (FA) respectively in IP-based wireless networks, their functionalities differ significantly. We focus on authentic NGWN, i.e., IP-based wireless

networks, and not only on mobility and interworking issues between 3G wireless networks.

A policy-enabled handoff decision algorithm proposed in Wang *et al.* (1999) is based on a cost function that considers several factors (e.g., bandwidth, power consumption and monetary costs). The cost function presented in Wang *et al.* (1999) is very preliminary and cannot handle more sophisticated scenarios. Also, cost function evaluation could require high processing time and power. In order to maximize user QoS, McNair & Zhu (2004) propose handoff decision algorithms for vertical handoff and identifies metrics that characterize NGWN. However, the proposed cost function could lead to singularity problems if connections become free of charge. Furthermore, handoff instability problem and mobility management at the IP layer are ignored. Many other vertical handoff schemes are presented in Zhu & McNair (2006). The factors considered in the above cited papers are insufficient. In fact, information about authentication types, access network types and roaming partners supported are not taken into account. Moreover, these studies do not provide a viable architecture framework for selection mechanisms, nor business models for prospective deployment.

4.3 Proposed Architecture for NGWN

A novel interworking architecture, called *Integrated InterSystem Architecture* (IISA) based on 3GPP/3GPP2-WLAN interworking models, is proposed and shown in Fig. 4.2. Instead of developing new infrastructures, IISA extends existing infrastructures to tackle integration and interworking issues and provides mobile users with ubiquity or *always best connected*. The IISA considers all of the above mentioned requirements (i.e., scalability, transparency, economics and security) for IP-based NGWN. Rather than adding an interworking entity between adjacent net-

works, as it is the case for some existing models presented in the literature such as BLR/BIU, IISA only adds a single new node, called *Interworking Decision Engine* (IDE) shown in Fig. 4.3, while other functionalities are implemented in the existing network components. Another main difference, between our approach and the BLR/BIU architecture, is the separation between the control plane (signaling traffic) and the transport plane (data traffic) in the IISA/HPIN proposal. In fact, only signaling traffic goes through the IDE, not data packets. In the BLR/BIU architecture, data packets and signaling traffic transit through BLR/BIU, thus creating bottlenecks in the system.

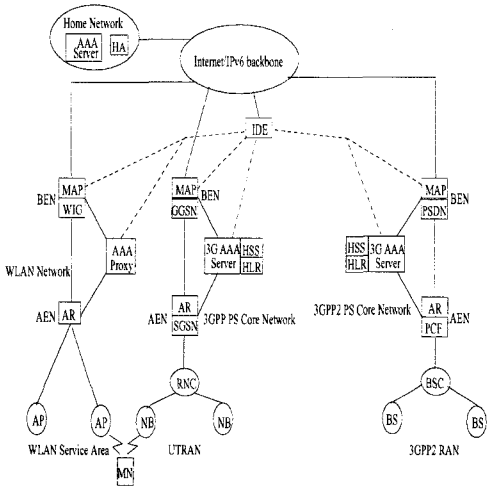

Figure 4.2 Integrated InterSystem Architecture (IISA).

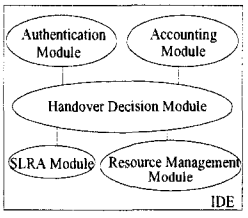

Figure 4.3 Interworking Decision Engine (IDE).

To enable support of IPv6-based mobility management protocols, some functional entities of 3G wireless networks are extended. The Serving GPRS (General Packet Radio Service) Support Node (SGSN) and Packet Control Function (PCF) are enhanced with the AR functionalities and are called *Access Edge Node* (AEN). Similarly, Gateway GPRS Support Node (GGSN) and Packet Data Serving Node (PDSN) are extended with MAP or HA and interworking functionalities (to enable

message format conversion, QoS requirement mapping, etc.) and are called *Border Edge Node* (BEN). The WLAN Interworking Gateway (WIG) acts as a route policy element, ensuring message format conversion. Extended functionalities can be integrated into existing networks entities or implemented separately. We advocate the first scenario as it is easily deployed and managed. The interworking of different access networks is required for an efficient integration. Mapping between HLR or home subscriber server (HSS) in 3G wireless networks and authentication, authorization and accounting (AAA) server/proxy in WLAN is required to execute authentication and billing when users roam across both technologies.

The IDE is introduced to enable interworking and handoff between various networks. Operators or service providers are required to have only one SLA with a third-party or IDE manager rather than establishing individual SLA with all of the other operators. The IDE allows reduction of signaling traffic, services disruption during handoff while handling AAA procedure and mobility management. To reduce the IDE's load, the IDE is involved only in intersystem and/or inter-domain handoff and it manages only control signaling traffic : data packet traffic bypass the IDE. Furthermore, to enable the scalability of the IISA architecture, if the number of mobile users that require intersystem and/or inter-domain handoff increases, or if the number of heterogeneous wireless systems increases, the IDE can be deployed within a hierarchical framework. For roaming users with sessions in progress, the IDE allows reduction of association and authentication delays. Usage of the IDE could be considered as a value-added service that network operators offer to their subscribers to allow roaming into other networks.

The *Authentication Module* (AuM) is used to authenticate users moving across different wireless networks and it avoids the required direct security agreements or associations between foreign networks and home network. The AuM stores information such as subscriber identity, user preferences, user profile and terminal mobility

patterns. The *Accounting Module* (AcM) enables billing between different wireless networks and stores charging information associated with the resources usage. It acts as common billing/charging system between various network operators. The AcM collects accounting information received from the AAA server/proxy of foreign network per user based on its billing policy. If necessary, it converts detailed call records of the foreign network before forwarding such information to the AAA server of the home network for billing purposes. The CIBER (Cellular Intercarrier Billing Exchange Roamer Record) protocol may be used for the exchange of roaming billing information among wireless operators through the IDE.

Usually different administrative domains have different QoS policies for resources allocation. Then, when an MN moves from one administrative domain to another, QoS re-negotiation may be required. Such re-negotiation will be based on SLAs between both domains. The *Resource Management Module* (RmM) enables the mapping of QoS parameters and their re-negotiation between various types of wireless networks and with the core IP transport network in order to satisfy the overall end-to-end QoS criteria. Furthermore, the RmM allows fast transfers of user profile and QoS requirements/parameters between two administrative domains during handoff. The QoS mapping and the mechanism by which the IDE allocates resources to an MN, and decides to admit a new request is outside the scope of this paper. However, we assume that the IDE is endowed with intelligence and can perform the following operations : translation of signaling message formats between different networks, conversion of higher transmission rate to lower rate, translation of QoS parameters and information, etc. The *SLRA Module* stores information about service providers or network operators that have SLAs and roaming agreements (RAs) with the IDE manager. The *Handover Decision Module* (HdM) is used when an intersystem or inter-domain handoff should be granted or not. In other words, it enables roaming and handoff support for MNs.

4.4 Proposed Handoff Protocol

Since mobility in NGWN is either logical or physical, user profile and preferences seem to be important when performing vertical handoff. As aforementioned, handoff decisions based on the RSS level are not appropriate in NGWN/4G. In Makaya & Pierre (2006), a handoff decision function is proposed for handoff decisions which take into account several parameters such as monetary cost, bandwidth, session priority, power consumption and network conditions, to enable efficient decisions and systems discovery. In the following sections, we propose a handoff management protocol that supports both vertical and horizontal handoffs in IPv6-based heterogeneous mobile environments. Under IISA, intra-MAP/BEN and inter-MAP/BEN roaming may result either into intrasystem or intersystem handoffs. Hence, HPIN is proposed for any of these types of handoff scenarios.

4.4.1 Authentication of Mobile Nodes

To avoid additional signaling overhead due to the execution of the AAA procedure each time an MN performs handoff and requests registration, we propose a token-based approach. While roaming within MAP/BEN domain of access networks having agreements with the IDE, an MN presents a token, that it obtains from the IDE (after its first successful registration in the visiting network) to the MAP/BEN or AR/AEN. The token includes security association parameters for secure tunnel setups between the MN and AR/AENs. This yields a lower registration latency than performing authentication and authorization check with AAA home server (AAAH). If the MAP/BEN or AR/AEN verifies the token successfully, it initiates an authorization process. The HA functionalities related to the MN authentication, distributing keying materials, session keys, security associa-

tion context and mobility management are delegated to the IDE while the MN roams in foreign networks. Subsequent authentications are handled either by the MAP/BEN and the AAA local server/proxy (AAAL) or by the IDE for intrasystem or intersystem movements.

4.4.2 Handoff Preparation with HPIN

With assumption that mobile devices will become increasingly powerful, intelligent and sensitive on changes of link layer, we adopt a network-assisted and mobile-controlled handoff strategy. The proposed handoff scheme combines mobile-monitored and network-probed information to provide reliable handoff control. Prior to handoff, an MN can obtain the information of candidate wireless networks to which it is likely to handoff, and uses this information to optimize the handoff performance. On the other hand, if mobile device capabilities are limited, handoff decisions are taken by mobility agents on the network side (e.g., IDE).

The MN decides whether to send the CARD Request message to the MAP/BEN according to generation of anticipated triggers (AT). For example, high bit error rate, link going down, weak signal strength, security risks, monetary cost and geographical location can be used as anticipated triggers. Upon generating anticipated triggers, the MN sends CARD Request message containing user preferences, applications required QoS capabilities to it serving MAP/BEN. With this message, the MN requests information of neighbor networks of its serving network to the IDE through the current MAP/BEN. With information exchanged between the MAP/BEN and candidate AR/AENs (CAR/AENs) by using the *Router Information eXchange* (RIX Request/Reply) messages, the MAP/BEN maintains a global view (i.e., load status of AR/AENs, connection state of any MN in its domain as well as movement patterns of all its serving MNs) of its domain and can learn both

link layer (L2) and IP layer (L3) information in an access network. Note that, if the CARD Request message was not sent in time, for example after generating the anticipated trigger and before generating the L2 trigger, HPIN will turn into HMIPv6. However, if the CARD Request message was not sent after generating the L2 trigger, HPIN turns into either FMIPv6 or F-HMIPv6.

L2 information may include the specific wireless access technology and the system parameters (e.g., channel frequency and number). On the other hand, L3 information may include the AR/AEN global address, the prefix of address advertised in wireless networks, the current QoS status and parameters. The QoS status include bandwidth availability and signal strength while QoS parameters may include information such as supported data rate, video coding rate and maximal delays. L2 and L3 information are then forwarded to the IDE and allows it to maintain a global view of all MAP/BEN domains having SLAs with the IDE manager. To allow seamless services continuity, requirements specified in the CARD Request message need to be setup consistently with the QoS negotiated in the previous subnet/subsystem. The QoS consistency is a highly challenging and crucial issue for real-time applications. This consistency is handled by the IDE, which allows QoS mapping between various networks. With this information, pre-filtering is performed by the MAP/BEN, based on the MN's preferences, the application required capabilities, network availability and the CAR/AEN list is obtained. If the MAP/BEN lacks user profile information, it requests such information to the IDE rather than to the MN's HA, which is usually far away from the current MAP/BEN.

The MAP/BEN responds to the MN through a CARD Reply message which contains the list of CAR/AEN. Upon receiving the CARD Reply message, the MN configures new on-link CoAs (NLCoAs) based on stateless IPv6 address autoconfiguration mode (Thomson & Narten, 1998). The MN can then start handoff at any time. CARD Request and Reply messages exchange do no longer delay the handoff

procedure, as it is performed while the MN uses the previous on-link CoA (PL-CoA). Whenever the L2 trigger is generated, using the information provided by the CARD Reply message, the MN can select which air-interface to turn on for access networks discovery and handoff preparation. L2 scanning process will be performed based on the information provided in the CAR/AEN list rather than scanning all frequencies or channels. Then, system discovery and L2 scanning process can be accelerated. This selective interface activation enables better tradeoff between system discovery time and power consumption efficiency compared to always on approach as used in most IPv6-based mobility management protocols. The MN will then compute the handoff decision function (Makaya & Pierre, 2006) for each reachable network contained in the CAR/AEN list, in order to determine whether there is a network with better QoS and select it as a target network.

4.4.3 Handoff Execution with HPIN

After the previous step, the MN sends a fast binding update (FBU) message to the serving MAP/BEN to notify the MAP/BEN that it is moving into new subnet/subsystem. Upon receiving FBU, the MAP/BEN starts a fast handoff procedure by sending a handoff initiate (HI) message to NAR/AEN, which includes a request to verify the pre-configured NLCoA and to establish a bi-directional tunnel between the MAP/BEN and NAR/AEN in order to prevent routing failure during handoff. In response to the HI message, the NAR/AEN performs a Duplicate Address Detection (DAD) procedure before responding with a handoff acknowledgment (HAck) message. After receiving the HAck message, the MAP/BEN sends the result to the MN by using a fast binding update acknowledgment (FBAck) message. Since the exact time when the MN will perform the link layer handoff is unpredictable, FBAck message is sent to both links, previous and new. This ensures that the MN receives the FBAck message either via the PAR/AEN or NAR/AEN

confirming the successful binding. Moreover, the MAP/BEN binds the PLCoA and NLCoA and it tunnels any packets addressed to PLCoA towards the NLCoA in the NAR/AEN's subnet. The NAR/AEN buffers these forwarded packets until the MN becomes attached to the NAR/AEN link.

The MN announces its presence on the new link by sending router solicitation (RS) with fast neighbor advertisement (FNA) option to the NAR/AEN. The FNA message is also used to confirm the usage of NLCoA when the MN has not received FBAck message through the previous link. Optionally, the NAR/AEN responds to the FNA message with a neighbor advertisement acknowledgment (NAAck) message to notify the MN to use another NLCoA, contained in FBAck rather than its prospective NLCoA, if there are addresses collision. Then, the NAR/AEN will start delivering buffered packets to the MN with FBAck most probably as the first packet on the new link. The bi-directional tunnel remains active until the MN completes the binding update procedure.

Note that, if the FBU message was not sent before the L2 handoff, then an MN sends it piggybacked in FNA message (FNA[FBU]) over the new link. When the NAR/AEN receives the FNA[FBU] message, it processes the FNA message part, extracts the FBU message part and forwards it to the serving MAP/BEN. When the serving MAP/BEN receives the FBU, it responds by sending FBAck message to NAR/AEN. At this time, the MAP/BEN can start tunneling towards NLCoA the incoming and in-flight packets addressed to PLCoA. This procedure refers to reactive mode of HPIN while the predictive mode is explained above (i.e., the MN sends the FBU through the PAR/AEN's link and the FBAck is received before the L2 handoff). The reactive mode can be carried either intentionally or serve as a fall-back solution when a predictive mode could not be completed successfully, for example, if the L2 handoff was completed before the FBAck message was received by the MN.

In case of inter-MAP/BEN roaming, the bi-directional tunnel is established between the previous MAP/BEN (MAP1/BEN in Fig. 4.5) and the NAR/AEN through the candidate MAP/BEN (MAP2/BEN in Fig. 4.5). Hence, the HI message is piggybacked in handoff request (HOReq) message and sent to the candidate MAP/BEN which processes the HOReq message part, extracts the HI message and forwards it to the target NAR/AEN. In response to the HI message, the NAR/AEN performs the DAD procedure before sending the HAck message. When the candidate MAP/BEN receives the HAck message, it includes this message in the handoff reply (HORep) message before forwarding it to the current MAP/BEN. After receiving HAck, the current MAP/BEN sends the result to the MN by using FBAck on both links (previous and new) and establishes binding between the previous and the new regional CoA (PRCoA and NRCoA), and tunnels any packets (buffered and incoming) addressed to PRCoA towards NRCoA. Message flow diagrams for both intrasystem or intersystem handoff during intra-MAP/BEN or inter-MAP/BEN roaming are illustrated in Fig. 4.4 and 4.5, respectively.

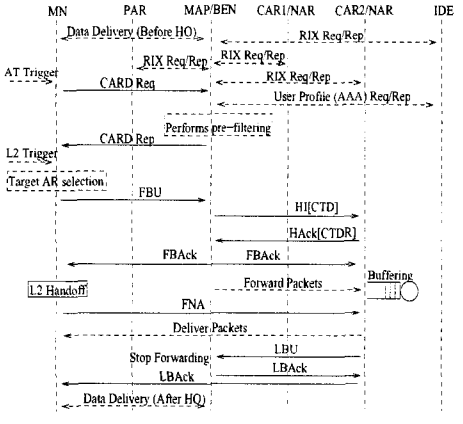

Figure 4.4 Signaling messages sequence in HPIN for intra-BEN roaming.

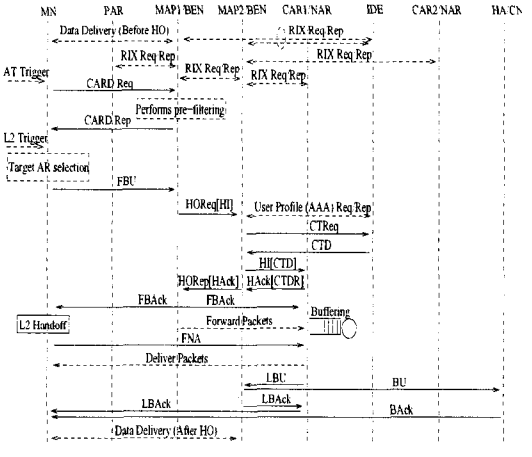

Figure 4.5 Signaling messages sequence with HPIN for inter-BEN roaming.

4.4.4 Context Transfer and Binding Update

Note that HI message triggers the request of context transfer rather than using Context Transfer Activate Request (CTAR) message as it is the case in the CXTP protocol (Loughney *et al.* , 2005). When the MAP/BEN receives a FBU message, it transmits a Context Transfer Data (CTD) message, piggybacked in HI, to the NAR/AEN containing feature contexts. Example of features contained in CTD message are QoS context information, header compression, security and AAA parameters. This paper mainly focuses on QoS context information. The routers extract this QoS context information, and according to context received, the intermediate router reserves corresponding resources and updates the path information. If the MAP/BEN has no context pertaining to the concerned MN, the new MAP/BEN sends a Context Transfer Request (CTReq) message to the IDE in order to obtain session management parameters for this MN and to establish traffic bearers on the new path. In response to a CTReq message, the IDE transmits a CTD message that includes the MN's previous IP address (i.e., RCoA) and feature contexts. When the MAP/BEN receives a CTD message, it installs the contexts as received from the IDE. The MAP/BEN includes the CTD message within the HI message and forwards it to the NAR/AEN.

When the NAR/AEN receives the CTD message, it may generate a CTD Reply (CTDR) message optionally to report the processing status of the received contexts and piggybacks this message in HAck. The NAR/AEN will send a HAck message to the MAP/BEN only after relocating traffic bearers and resources reservation (resource reservation procedure is out of the scope of this paper) towards the new path, in order to indicate that handoff may be conducted and packets forwarding may start. Hence, unlike to FMIPv6 and F-HMIPv6 where that forwarded packets have no QoS guarantee before the new QoS path is setup, HPIN solves this issue.

The binding update (BU) procedure is performed by the NAR/AEN on the behalf of MNs. In fact, an AR/AEN acts as a proxy : copies a BU list of an MN in its cache and manages this list (e.g., lifetime entries) in the same way as the original is managed by the MN. The AR/AEN cache copy must be updated periodically according to the original BU list of the MN. The BU list contains information about used home address and CoA (LCoA and RCoA), IPv6 address of CNs, sequence number, lifetimes, and state of retransmissions. When the BU list lifetimes cached in AR/AEN is about to expire, the AR/AEN may send a BU list renewal request to the MN. The BU list renewal is performed in the same way as a classical BU refresh (Johnson *et al.* , 2004). By piggybacking the BU list in a FNA message, separate out-of-band messages from MN to NAR/AEN are avoided, thus reducing signaling traffic overhead.

4.5 Analytical Model for HPIN

In IP-based wireless networks, QoS may be defined by packet loss, handoff latency, handoff blocking probability and signaling overhead. Analyses of these metrics are very useful in order to evaluate the performance of mobility management protocols. The notation used in this paper is given in Table 4.1. Let χ_T be the random variable for the time between the L2 trigger generation and the link down (i.e., pending L2 handoff) and let $f_T(u, \sigma)$ be the probability density function for successful completion of signaling, where $\sigma > 0$ is a success rate parameter. The probability P_s of anticipated handoff signaling success for a particular observed valued t_T is expressed by :

$$P_s = Pr(\chi_T > t_T) = \int_{t_T}^{\infty} f_T(u, \sigma) du. \tag{4.1}$$

Tableau 4.1 Notation.

t_s	inter-session time between two consecutive sessions with PDF f_s
t_c	subnet (AR/AEN's coverage area) residence time with PDF f_c
t_d	MAP/BEN domain residence time with PDF f_d
N_c	number of subnet crossings during intra-MAP/BEN roaming
N_d	number of MAP/BEN domain crossings during inter-BEN roaming
C^g	global binding update cost to HA/CNs
C^l	local binding update cost to MAP/BEN
M	number of subnets in MAP/BEN domain
N_{CN}	number of CNs with a binding cache entry for an MN
$d_{X,Y}$	average number of hops between nodes X and Y
κ(resp. τ)	unit transmission cost over wired (resp. wireless) link
$C_{X,Y}$	transmission cost of control packets between nodes X and Y
PC_X	processing cost for binding update at node X
t_T	time period between the L2 trigger and the start of the link switching

Deriving an expression for P_s is difficult, since it depends on the exact form of $f_T(u, \sigma)$, which is usually unknown. For the sake of simplicity, we assume that χ_T is exponentially distributed.

4.5.1 User Mobility and Traffic Models

User mobility and traffic models are crucial for efficient system design and performance evaluation. Usually, an MN mobility is modeled by the cell residence time and various random variables type are used for this purpose (Fang, 2003). Two commonly used mobility models in wireless networks are : random-walk and fluid-flow models (Wang & Akyildiz, 2000). Evaluating the time span that an MN will stay within the subnet is usually based on two distributions : exponential and Gamma. The Gamma distribution is very realistic for mobility models as it considers changes in the MN speed and direction.

On the other hand, although the incoming calls or sessions in NGWN follow a Poisson process (i.e., inter-arrival time are exponentially distributed), the inter-service time is not necessarily exponentially distributed (Fang, 2003). Other distribution models, such as hyper-Erlang and Pareto, have been proposed. Furthermore, the self-similar nature of data traffic has been noticed. However, performance evaluations reported in the literature show that the exponential model can be appropriate for cost analyses. In fact, the exponential model provides an acceptable tradeoff between complexity and accuracy. Hence, most cost analyses adopt exponential assumption (Fang, 2003). We consider a traffic model composed of two levels, session and packets. The session duration follows an exponential distribution with inter-session rate λ_s while packet generation follows a Poisson process.

Let μ_c and μ_d be the border crossing rate for an MN out of a subnet (i.e., AR/AEN domain) and a MAP/BEN domain, respectively. When an MN crosses a MAP/BEN domain border, it also crosses an AR/AEN border. Then, let μ_l be the border crossing rate for which an MN still stays in same MAP/BEN domain, $\mu_l = \mu_c - \mu_d$. Under the fluid-flow mobility model, let v represents the average velocity of an MN, ρ the user density and L_c express the perimeter of a subnet. The subnet crossing rate can be computed by : $\mu_c = \frac{\rho v L_c}{\pi}$. If we assume that all subnets have a circular shape and form together a contiguous area and that each MAP/BEN domain is composed of M equally subnets, we obtain : $\mu_d = \frac{\mu_c}{\sqrt{M}}$.

Modeling the probability distribution of the number of boundary crossings during a session lifetime plays a significant role in cellular networks cost analyses. The same will apply to IP-based wireless networks. For the sake of simplicity and to derive analytical expressions easily, the exponential distribution will be used. The roaming probability depends on an MN's movement pattern in its original network but not in its destination network. Hence, the probability that there is at least one local (resp. global) binding update between two consecutive sessions of an MN, P_c

(resp. P_d) is expressed by :

$$\begin{aligned} P_c &= Pr(t_s > t_c) = \int_0^\infty Pr(t_s > u) f_c(u) du = \frac{\mu_c}{\mu_c + \lambda_s} \\ P_d &= Pr(t_s > t_d) = \int_0^\infty Pr(t_s > u) f_d(u) du = \frac{\mu_d}{\mu_d + \lambda_s}. \end{aligned} \tag{4.2}$$

Probabilities that an MN experiences k subnets boundary crossings and m access network boundary crossings during the lifetime of its session correspond to probabilities mass function (PMF) of random variables N_c and N_d, respectively and are expressed as follows (Xiao *et al.* , 2004) :

$$Pr(N_c = k) = P_c^k(1 - P_c) \quad \text{and} \quad Pr(N_d = m) = P_d^m(1 - P_d). \tag{4.3}$$

Then, the average number of location binding updates during an inter-session time interval under subnet crossings, $E(N_c)$, and MAP/BEN domain crossings, $E(N_d)$, are given by :

$$\begin{aligned} E(N_c) &= \sum_{k=0}^{\infty} kPr(N_c = k) = \sum_{k=0}^{\infty} kP_c^k(1 - P_c) = \frac{\mu_c}{\lambda_s} \\ E(N_d) &= \sum_{m=0}^{\infty} mPr(N_d = m) = \sum_{m=0}^{\infty} mP_d^m(1 - P_d) = \frac{\mu_d}{\lambda_s}. \end{aligned} \tag{4.4}$$

With the same assumption on time variables, we can obtain the expression of $E(N_l)$, i.e., the average number of subnets that an MN crosses and still stay within a given MAP/BEN domain during an inter-session time interval, as follows : $E(N_l) = \mu_l/\lambda_s$.

4.5.2 Total Signaling Cost

Performance analyses of wireless networks must consider a total signaling cost induced by a mobility management scheme. As for wireless cellular networks, signaling traffic overhead cost must be computed for NGWN or IP-based mobile

environments. NGWN supports two kinds of location or binding updates. One occurs from an MN's subnet boundary crossing and the other occurs when the binding lifetime is about to expire. Moreover, data packet delivery induces usage of network resources, thus generating an additional cost. Hence, the total signaling cost (C_T) could be considered as the sum of binding update signaling cost (C_{BU}) and packet delivery cost (C_{PD}), and given by : $C_T = C_{BU} + C_{PD}$.

4.5.2.1 Binding Update Signaling Cost

Depending on the movement type, two kinds of binding update can be performed : *local* and *global*. The global binding update occurs when an MN moves out of its MAP/BEN domain. In this case, the MN registers its new regional CoA (RCoA) to HA and to active CNs. On the other hand, if the MN changes its current address (LCoA) within a MAP/BEN domain, it only needs to register this new LCoA to the MAP/BEN. Hence, the average binding update signaling cost during inter-session time intervals heavily depends on computed number of binding updates :

$$C_{BU} = E(N_l)C^l + E(N_d)C^g. \tag{4.5}$$

To perform signaling overhead analyses, a performance factor called session-to-mobility ratio (SMR) is introduced. It is similar to the call-to-mobility ratio (CMR) defined in wireless cellular networks (Xie & Akyildiz, 2002). The SMR represents the relative ratio of session arrival rate over user mobility rate : $SMR = \lambda_s/\mu_c$. The binding update signaling cost, C_{BU}, is then given by :

$$C_{BU} = \frac{1}{\lambda_s}\left(\mu_d C^g + \mu_l C^l\right) = \frac{1}{SMR\sqrt{M}}\left[C^g + (\sqrt{M}-1)C^l\right]. \tag{4.6}$$

Anticipated trigger and link layer information (L2 trigger) are used either to

Tableau 4.2 Expression of signaling costs.

$C_{mu} = 2C_{NAR,BEN} + PC_{BEN}$
$C_{ru} = 2(C_{BEN,HA} + N_{CN}C_{BEN,CN}) + PC_{HA} + N_{CN}PC_{CN}$
$S_f^l = C_{MN,BEN} + 3C_{BEN,NAR} + 2PC_{BEN} + PC_{NAR}$
$S_s^l = 2C_{MN,BEN} + 3C_{BEN,NAR} + C_{MN,NAR} + 2PC_{BEN} + 2PC_{NAR}$
$S_r^l = C_{MN,NAR} + 2C_{NAR,BEN} + PC_{BEN} + 2PC_{NAR}$
$S_f^g = C_{MN,pBEN} + 3(C_{pBEN,nBEN} + C_{nBEN,NAR}) + 4PC_{BEN} + PC_{NAR}$
$S_r^g = C_{MN,NAR} + 2(C_{NAR,nBEN} + C_{nBEN,pBEN}) + 2PC_{BEN} + PC_{NAR}$
$S_s^g = 2C_{MN,pBEN} + 3(C_{pBEN,nBEN} + C_{nBEN,NAR}) + C_{MN,NAR} + 4PC_{BEN} + 2PC_{NAR}$

predict or rapidly respond to handoff events. Hence, HPIN signaling cost depends on the probability that handoff anticipation is relevant or not. The critical phase of the HPIN starts when the L2 trigger is generated and indicates the imminence of the handoff. We assume that if an MN receives a FBAck message from the MAP/BEN, that it will definitely start the L3 handoff to NAR/AEN without exceptions. Hence, if there is no real handoff after L2 trigger, all messages exchanged for handoff anticipation may be unnecessary. Thus, global and local binding update signaling costs for HPIN are expressed as follows :

$$\begin{aligned} C^g &= P_s S_s^g + (1 - P_s)(S_f^g + S_r^g) + C_{ru} \\ C^l &= P_s S_s^l + (1 - P_s)(S_f^l + S_r^l) + C_{mu} \end{aligned} \tag{4.7}$$

where C_{ru} represents the binding update cost at the IDE or at HA/CNs; C_{mu} depicts the binding update cost at MAP/BEN; S_s^g (resp. S_s^l) denotes the global (resp. local) signaling cost for successfully anticipated handoff; S_f^g (resp. S_f^l) the global (resp. local) signaling cost for control messages bear if no real L3 handoff occurs and S_r^g (resp. S_r^l) indicates the global (resp. local) signaling cost for the HPIN reactive mode. Expression of those signaling costs are given in Table 4.2.

4.5.2.2 Packet Delivery Cost

Similarly to investigation reported in Koodli & Perkins (2001), handoff latency is subdivided into three components : (1) link switching or L2 handoff latency, t_{L2}, (2) IP connectivity latency (t_{IP}) due to movement detection and address configuration and (3) location update latency (t_U). The IP connectivity latency reflects how quickly an MN can send IP packets after L2 handoff, while location update latency is the delay required to forward IP packets to the MN's new IP address. On the other hand, the time period from the starting point of L2 handoff to when an MN receives IP packets for the first time after link switching refers to packet reception latency (t_P) or data latency. Moreover, the following delay components are defined : binding update latency (t_{BU}) and delay from completion of binding update and reception of the first packet by an MN through the new IP address (t_{NR}).

When two endpoints have an ongoing session, a packet delivery cost incurs. The packet delivery cost is composed of packet transmission and packet processing costs. By using the handoff timing diagram illustrated in Fig. 4.6, the packet delivery cost could be defined as the linear combination of packet tunneling cost (C_{tun}) and packet loss cost (C_{loss}). Let α and β be the weighting factors which emphasize the tunneling and dropping effect. The packet delivery cost, C_{PD}, is computed as follows :

$$C_{PD} = \alpha C_{tun} + \beta C_{loss}. \tag{4.8}$$

To avoid packet loss, HPIN enables the MAP/BEN to forward packets to NAR/AEN by using a tunnel established between them and the NAR/AEN buffers all forwarded packets. The HPIN timing diagram for intra-MAP/BEN movement is shown in Fig. 4.6.

In IP networks, the signaling cost is proportional to the distance in hops between

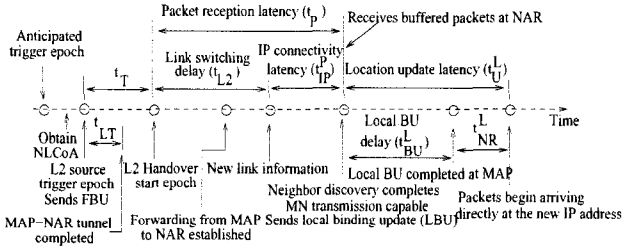

Figure 4.6 Timing diagram of HPIN for intra-BEN roaming.

the source and destination nodes. Furthermore, the transmission cost in a wireless link is generally larger than the transmission cost in a wired link (Xie & Akyildiz, 2002). Let s_c and s_d be the average size of control packets and data packets, respectively and $\eta = s_d/s_c$. The cost of transferring a data packet is η greater than the cost of transferring a control packet. Let λ_p be the packet arrival rate in unit of packet per time. The packet tunneling cost for HPIN predictive mode can be expressed as follows :

$$C_{tun}^{p,l} = \lambda_p C_{cm}^{s,l}(t_{L2} + t_{IP}^P + t_U^L) \tag{4.9}$$

where $C_{cm}^{s,l} = \eta(C_{CN,BEN} + C_{BEN,NAR} + C_{NAR,MN})$ is the cost of transferring data packets from the CN to an MN by transiting to MAP/BEN ; t_U^L denotes the location update latency for intra-MAP/BEN movement ($t_U^L = t_{BU}^L + t_{NR}^L$) and t_{IP}^P depicts the IP connectivity latency excluding the IP address configuration, DAD procedure and movement detection. In fact, these operations are conducted before an MN leaves the PAR/AEN's link.

In fast handoff schemes, packets loss are due either to L2 handoffs or when an MN moves to another subnet before a forwarding tunnel has been established. The latter case refers to wrong temporal and spatial predictions. Packet loss due to L2 handoff delay is inevitable without efficient buffering mechanisms (Koodli & Perkins, 2001). Let t_{LT}, the time required to establish a tunnel between the MAP/BEN and the NAR/AEN. Usually, t_T is greater than t_{LT}, thus, packets received during

handoff are forwarded to NAR/AEN by the MAP/BEN using the already established tunnel. However, if the MN moves very fast, t_T may be inferior to t_{LT}. Then, packets arriving to MAP/BEN during the time period $t_{LT} - t_T$ may be lost, since the tunnel is not yet established. In other words, for the anticipated signaling to succeed, the following time constraint must be observed : $t_{LT} \leq t_T$. Hence, the cost associated with the packet loss can be expressed as follows :

$$C_{loss}^{p,l} = \lambda_p C_{cm}^{f,l} \max(t_{LT} - t_T, 0) \tag{4.10}$$

where $C_{cm}^{f,l} = \eta(C_{CN,BEN} + C_{BEN,PAR} + C_{PAR,MN})$ is the cost of transferring data packet from the CN to the MN by transiting to MAP/BEN when handoff fails or if the binding update is not yet performed at the MAP/BEN.

Due to wrong spatial predictions of NAR/AEN, or if a FBAck message was not received through the previous link, packets forwarded to a mispredicted NAR/AEN by the MAP/BEN may be lost. The process of forwarding packets to the wrong NAR/AEN is stopped when the FBU message sent through NAR/AEN's link is received at the MAP/BEN. Moreover, if an MN's movement within subnet overlapping area is longer than the tunnel establishment delay, the HPIN turns into its reactive mode. Since the packet forwarding process is not supported in the reactive mode, the packet tunneling cost equals zero ($C_{tun}^{r,l} = 0$) while the HPIN packet loss cost can be expressed as follows :

$$C_{loss}^{r,l} = \lambda_p C_{cm}^{f,l}(t_{L2} + t_{IP}^{R} + t_{U}^{L}) \tag{4.11}$$

where t_{IP}^{R} is the IP connectivity latency of reactive mode for an intra-MAP/BEN movement. The average packet delivery cost of HPIN scheme is then given by :

$$C_{PD}^{a,l} = P_s C_{PD}^{p,l} + (1 - P_s) C_{PD}^{r,l} \tag{4.12}$$

where $C_{PD}^{p,l}$ and $C_{PD}^{r,l}$ are packet delivery costs for the HPIN predictive and reactive mode and are computed by (4.8).

The timing diagram of HPIN for inter-MAP/BEN roaming is illustrated in Fig. 4.7. With similar reasoning as for intra-MAP/BEN, packet tunneling cost (C_{tun})

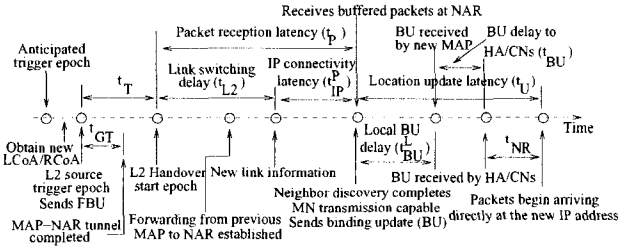

Figure 4.7 Timing diagram of HPIN for inter-BEN roaming.

and packet loss cost (C_{loss}) for inter-MAP/BEN roaming with HPIN, are expressed as follows :

$$\begin{aligned} C_{tun}^{p,g} &= \lambda_p C_{cm}^{s,g}(t_{L2} + t_{IP}^P + t_U) \\ C_{loss}^{p,g} &= \lambda_p C_{cm}^{f,g} \max(t_{GT} - t_T, 0) \\ C_{loss}^{r,g} &= \lambda_p C_{cm}^{f,g}(t_{L2} + t_{IP}^{RG} + t_U) \end{aligned} \tag{4.13}$$

where for inter-MAP/BEN roaming, $t_U = t_{BU} + t_{RR} + t_{NR}$, t_{RR} is the delay to complete the return routability procedure, t_{IP}^{RG} is the IP connectivity latency for the reactive mode, t_{GT} is the time required to establish a tunnel between the previous MAP/BEN and the NAR/AEN, $C_{cm}^{f,g} = \eta(C_{CN,pBEN} + C_{pBEN,PAR} + C_{PAR,MN})$ and $C_{cm}^{s,g} = \eta(C_{CN,pBEN} + C_{pBEN,nBEN} + C_{nBEN,NAR} + C_{NAR,MN})$.

4.5.3 Handoff Latency and Packet Loss

The following parameters are defined to compute handoff latency and packet loss : t_{L2} indicates the L2 handoff latency and $t_{X,Y}$ specifies one-way transmission delay between nodes X and Y for a message of size s. If one of the endpoints is an

MN, $t_{X,Y}$ is computed as follows :

$$t_{X,Y}(s) = \frac{1-q}{1+q}\left(\frac{s}{B_{wl}} + L_{wl}\right) + (d_{X,Y} - 1)\left(\frac{s}{B_w} + L_w + \varpi_q\right) \tag{4.14}$$

where q is the wireless link failure probability, ϖ_q is the average queueing delay at each router in the Internet (McNair *et al.* , 2001), B_{wl} (resp. B_w) denotes the wireless (resp. wired) link bandwidth and L_{wl} (resp. L_w) expresses wireless (resp. wired) link delay.

Let Δ_{ns} be the time elapsed between receiving the FBAck on the previous link and the beginning of the L2 handoff when L2 and L3 handoff operations are not well synchronized. Moreover, let Δ_{lr} be the time between last packet reception through the previous link and L2 handoff beginning when FBAck is received on the new link. Note that Δ_{lr} and Δ_{ns} may be equal zero. For HPIN, the handoff latency depends on the information available, and on which link fast handoff messages are exchanged. If information about NAR/AEN and an impending handoff are available, and if FBAck message is received through the previous link, the handoff latency is expressed as follows :

$$O^l_{HPIN} = \Delta_{ns} + t_{L2} + 2t_{MN,NAR}. \tag{4.15}$$

However, if a FBAck message is not received through the previous link, it will be received through the new link. In this case the handoff latency for HPIN is expressed as follows :

$$N^l_{HPIN} = \Delta_{lr} + t_{L2} + 2t_{MN,NAR} + 3t_{NAR,BEN}. \tag{4.16}$$

The average handoff latency with HPIN for intra-MAP/BEN roaming is given as

follows :

$$D^l_{HPIN} = P_s O^l_{HPIN} + (1 - P_s) N^l_{HPIN}. \tag{4.17}$$

For inter-MAP/BEN roaming case, when the FBAck message is received through the previous link, the handoff latency associated to HPIN is identical to intra-MAP/BEN roaming : $O^g_{HPIN} = O^l_{HPIN}$. In fact, the handoff procedure depends only on intra-MAP/BEN communication delay, since the inter-MAP/BEN signaling is completed before the L2 handoff. On the other hand, when the FBAck message is received through the new link for inter-MAP/BEN movement, we assume that appropriates information about the NAR/AEN are already available and NLCoA is already configured. Hence, the handoff latency with HPIN for inter-MAP/BEN roaming is given by :

$$N^g_{HPIN} = \Delta_{lr} + t_{L2} + 2t_{MN,NAR} + 3[t_{NAR,nBEN} + t_{nBEN,pBEN}]. \tag{4.18}$$

The average HPIN handoff latency for inter-MAP/BEN roaming is computed similarly as in (4.17). With HPIN, in theory, no packets are lost, unless buffers overflow at NAR/AEN or MAP/BEN. However, without efficient buffer management, forwarded packets can be lost during handoff latency. In fact, the number of packets lost is proportional to handoff latency.

4.5.4 Handoff Blocking Probability

The handoff blocking probability is used to express the likelihood that a session/call connection will be terminated prematurely due to unsuccessful handoff during a session lifetime. Subscribers are more sensitive to session blocking during handoff than at the moment the call is initiated. Hence, minimizing the handoff blocking probability is crucial for mobility management schemes. The handoff blo-

cking can be caused by many factors, including handoff latency, signal-to-noise ratio (SNR) deterioration, unavailable channel, session rejection by the target network. However, this analysis considers only latency as a handoff blocking factor.

When an MN moves from one subnet to another, if the subnet residence time is less than the total handoff time, packets are lost and service is forcefully terminated due to loss of link or channel. Let T_S be the random variable defining the signaling delay due to the handoff and $\widetilde{T}_S$ the mean value of the total handoff latency. If we assume that T_S is exponentially distributed with cumulative density function $F_T(t)$, the handoff blocking probability is given by :

$$P_B = Pr(T_S > t_c) = \int_0^\infty [1 - F_T(u)] f_c(u) du = \frac{\mu_c \widetilde{T}_S}{1 + \mu_c \widetilde{T}_S}. \tag{4.19}$$

4.5.5 Processing Load of the IDE

Wireless overlay networks are subdivided into low-tier (e.g., WLAN) and high-tier (e.g., 3G wireless network) (Stemm & Katz, 1998). Roaming between low-tier and high-tier networks refers to vertical or intersystem handoff. To analyze the load incurred at the IDE, we assume that high-tier networks fully overlap low-tier networks and users are uniformly distributed. Let N_h and N_l be the number of high-tier and low-tier networks in the service or coverage area (e.g., one city), respectively. User density is denoted by ρ_h in high-tier and ρ_l in low-tier networks.

Recall that with MIPv6, each subnet crossing results in a binding update to the HA. Moreover, during refresh time period, each MN sends out a refresh request to the HA. Thus, the processing load at the HA with MIPv6 scheme is :

$$L_{HA} = P_{BU} \frac{[N_l \rho_l v_l L_l + N_s \rho_h v_h L_s]}{\pi} + P_{BR} \frac{[\nu_l \rho_l A_l N_l + \nu_h \rho_h A_h N_h]}{T_{HA}} \tag{4.20}$$

where N_s is the total number of subnets in a high-tier network, $N_h \leq N_s$, ν_l (resp.

ν_h) stands for the proportion of subscribers in low-tier (resp. high-tier) network away from their home network, P_{BU} is the processing time for an update registration message and P_{BR} depicts the processing time for binding refresh message. T_{HA} and T_{IDE} denote the binding lifetime period at the HA and the IDE, while A_l and A_h indicate the coverage area of low-tier and high-tier networks. On the other hand, v_l and v_h are the average speed of an MN in low-tier and high-tier networks, L_l is the perimeter of low-tier network while L_s is for a subnet in high-tier network.

In HPIN, binding refresh and binding update are performed locally at the MAP/BEN and not to the IDE as long as an MN moves within the MAP/BEN domain or performs intrasystem handoffs. However, during the refresh time period T_{IDE} the MAP/BEN sends one RIX (Request or Reply) message to the IDE for a given number of MNs. We denote ε_l the number of these MNs for low-tier networks and ε_h for high-tier networks. Therefore, when intersystem and/or inter-domain handoff occurs, path updates are required. Thus, the IDE processing load is expressed by :

$$L_{IDE} = P_{PU}\frac{[N_l\rho_l v_l L_l + N_h\rho_h v_h L_s]}{\pi} + P_{PR}\frac{\left\lceil\frac{\nu_l\rho_l A_l N_l}{\varepsilon_l}\right\rceil + \left\lceil\frac{\nu_h\rho_h A_h N_h}{\varepsilon_h}\right\rceil}{T_{IDE}} \tag{4.21}$$

where P_{PU} stands for the processing time for path updates and P_{PR} is the processing time for path refresh message. Comparing (4.20) and (4.21) clearly shows that $L_{IDE} \leq L_{HA}$. On the other hand, assume that there are O operators in the service area. The number of bilateral SLAs required to realize a roaming among all networks deployed with traditional interworking architecture is $\frac{O(O-1)}{2}$. The number of SLAs required with the IISA architecture is O. When O is very high, IISA allows a significant reduction on the number of SLAs.

Tableau 4.3 Performance analysis parameters.

Parameters	Symbols	Values
L2 handoff time	t_{L2}	50 msec
Time period between L2 trigger and L2 handoff	t_T	10 msec
Prediction probability	P_s	0.98
Wireless link failure probability	q	0.50
Wired link bandwidth	B_w	100 Mbps
Wireless link bandwidth	B_{wl}	11 Mbps
Wired link delay	L_w	2 msec
Wireless link delay	L_{wl}	10 msec
Control packet size	s_c	96 bytes
Data packet size	s_d	200 bytes
Packet arrival rate	λ_p	10 packets/s
MN average speed	v_l, v_h	5.6 Km/h
Low-tier subnet radius	R_l	50 m
High-tier subnet radius	R_s	1000 m
User density in high/lower-tier networks	ρ_h, ρ_l	0.002 m^{-2}

4.6 Performance Evaluation

An analytical framework to evaluate the performance of IPv6-based handoff schemes proposed by the IETF (i.e., MIPv6, HMIPv6, FMIPv6 and F-HMIPv6) is presented in Makaya & Pierre (2007a). Such evaluation methods will be used to compare the performance of the IETF's protocols and with HPIN. The parameter values used in the performance evaluation are given in Table 4.3, except when wireless link delay (L_{wl}), packet arrival rate (λ_p), prediction probability (P_s) and user density in low-tier networks (ρ_l) are considered variable parameters.

The network topology considered for analysis is illustrated in Fig. 4.8. We assume the distance between different domains to be equal, i.e., $c = d = e = f = 10$ and set $a = 1$, $b = 2$, and $g = 4$. All links are supposed to be full-duplex in terms of capacity and delay. Parameter values used to compute signaling cost are defined as

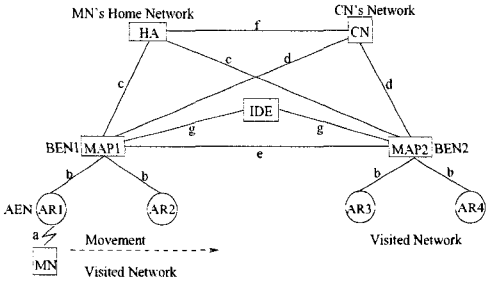

Figure 4.8 Network topology used for analysis.

follows : $\tau = 1$, $\kappa = 10$, $\alpha = 0.2$, $\beta = 0.8$, $PC_{AEN} = 8$, $PC_{HA} = 24$, $PC_{CN} = 4$, $PC_{IDE} = 15$ and $PC_{BEN} = 12$. The values of other parameters are : $\varepsilon_l = \varepsilon_h = 10$, $N_l = 40$, $N_h = 5$, $N_s = 15$, $\nu_l = \nu_h = 0.1$, $T_{HA} = T_{IDE} = 20$ min, $P_{BU} = 0.008$ msec, $P_{BR} = 0.001$ msec, $P_{PU} = 0.002$ msec, and $P_{PR} = 0.005$ msec.

Fig. 4.9 illustrates the binding update signaling cost as a function of the SMR. When the SMR is small, the mobility rate is superior to the session arrival rate, the MN frequently changes its point of attachment, resulting in several handoffs. Then, the signaling traffic overhead increases. The signaling overhead is considerably reduced from FMIPv6 to HPIN. However, when the session arrival rate is greater than the mobility rate (i.e., $SMR > 1$), the binding update is performed less often. In other words, signaling overhead decreases as the subnet change frequency decreases. The HPIN enables significant cost saving in terms of signaling overhead. Additional messages introduced in HPIN to allow handoff anticipation cause the signaling overhead to increase slightly compared to HMIPv6. However, this signaling overhead increment is compensated by lower handoff latency and packet loss as shown below. The packet delivery cost is depicted in Fig. 4.10 as a function of packet arrival rate (λ_p). HPIN outperforms all other IPv6-based handoff management schemes and HPIN is more efficient when λ_p increases. This means that HPIN is highly suitable for real-time applications where periodic packets are sent at high rate.

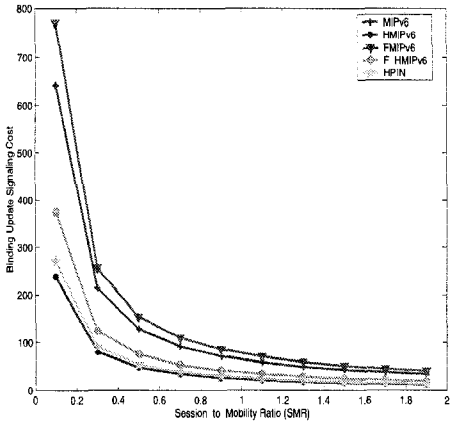

Figure 4.9 Binding update signaling cost.

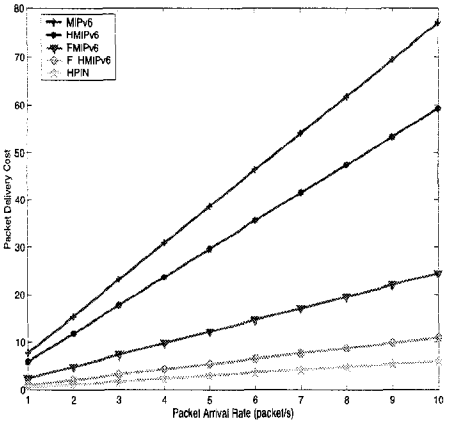

Figure 4.10 Packet delivery cost vs. packet arrival rate.

Fig. 4.11 shows the packet delivery cost for varying prediction probability (P_s). The packet delivery cost decreases as the accuracy of P_s increases in fast handoff schemes. Higher P_s value means that the FBAck message is received through the previous link. Then, packets are delivered to an MN just after being attached to the NAR/AEN. Results show that HPIN performs better than all other schemes as it provides a lower packet delivery cost. The prediction probability has a huge effect on F-HMIPv6 and if $P_s = 0$, F-HMIPv6 turns into HMIPv6, its reactive mode.

Fig. 4.12, shows that the handoff latency increases linearly with the wireless link delay. MIPv6 has a worst performance compared to other schemes, followed by HMIPv6. Furthermore, FMIPv6 and F-HMIPv6 allow handoff latency reduction for MIPv6 and HMIPv6. Also, HPIN allows a significant handoff latency reduction compared to other mobility management protocols. It is well known that the maximal tolerable delay for interactive conversation is approximately 200 msec. Hence, HPIN meets this requirement when the wireless link delay is set below 60 msec. The effect of prediction probability (P_s) on handoff latency is shown in Fig. 4.13. Regardless of the P_s value, HPIN performs better than all the other protocols.

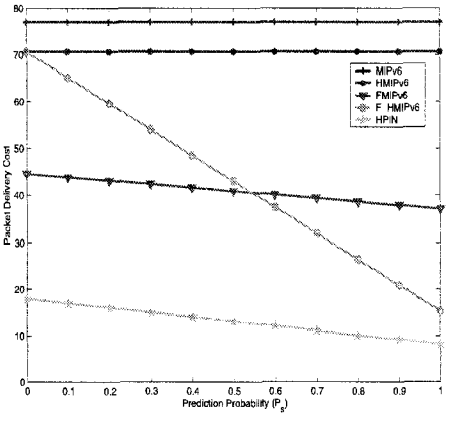

Figure 4.11 Packet delivery cost vs. prediction probability.

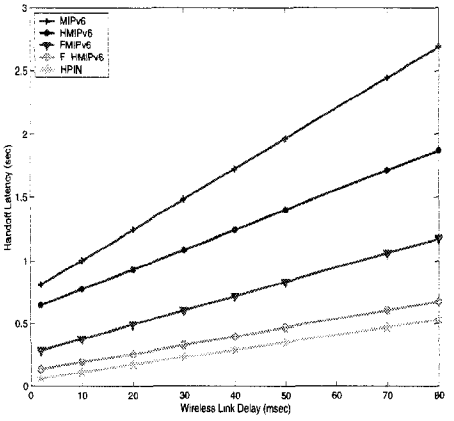

Figure 4.12 Handoff latency vs. wireless link delay.

Fig. 4.14 shows the total packet loss in terms of packet arrival rate. Packet loss values are much lower for HPIN than other IPv6-based handoff protocols. The effect of handoff in IPv6-based wireless environments is dominated by packet loss, which is due to L2 handoff and the IP layer operations. In fact, due to the lack of buffering and anticipated handoff mechanisms in MIPv6 and HMIPv6, all in-flight packets are lost during handoff. However, in fast handoff schemes (i.e., FMIPv6, F-HMIPv6 and HPIN) packet loss begins when L2 handoff is detected until the buffering mechanism is initiated or if buffers overflow. Fig. 4.15 shows that HPIN has much lower handoff blocking probability than other IPv6-based handoff schemes. This result is due to the ability of HPIN to reduce signal message exchanges and handoff latency. Thus, HPIN can safely provide seamless handoff with services continuity.

Fig. 4.16 shows the impact of the number of low-tier networks on the processing load for different values of the MN's average speed. Results show that the IDE processing load is lower than at the HA required for MIPv6. Thus, the IDE load due to intersystem and/or inter-domain handoffs is limited. On the other hand, one HA is usually used to handle MIPv6 handoff in service coverage area (e.g., one city)

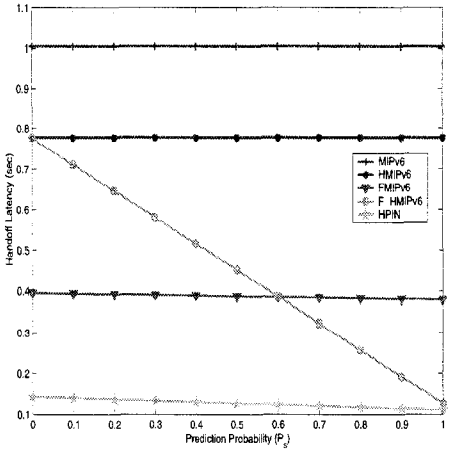

Figure 4.13 Handoff latency vs. prediction probability.

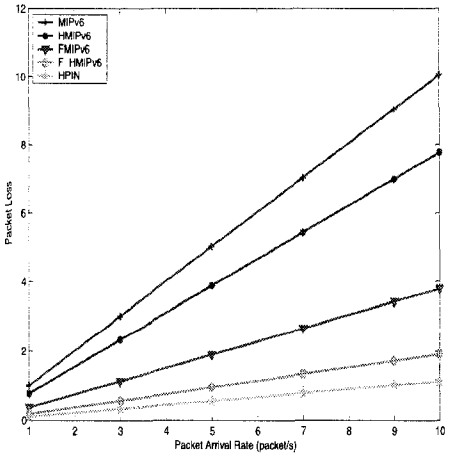

Figure 4.14 Packet loss vs. packet arrival rate.

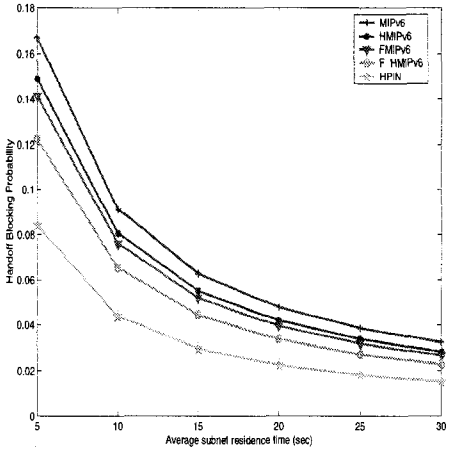

Figure 4.15 Comparison of handoff blocking probability.

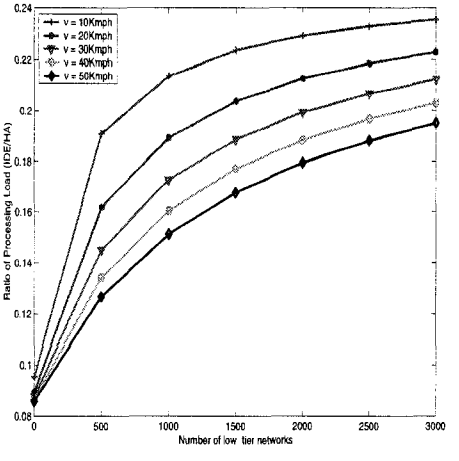

Figure 4.16 Processing load ratio vs. number of low-tier networks.

by network operators. We can thus conclude that a single IDE will be sufficient to handle intersystem and/or inter-domain handoffs for a coverage area of one city.

Fig. 4.17 illustrates that, as user density increases, the processing load for inter-system and/or inter-domain handoffs at the IDE remains insignificant compared to the processing load at the HA for MIPv6. Fig. 4.18 shows that the IDE processing load increases as the number of cities increases. This means that the IDE load increases proportionally to the size of the service coverage area. Therefore, an MN with a higher average velocity is associated with a greater domain crossing rate, which results into a higher number of handoff requests. Such results encourage the deployment of the IDE through hierarchical architecture to allow the integration and the interworking of various networks.

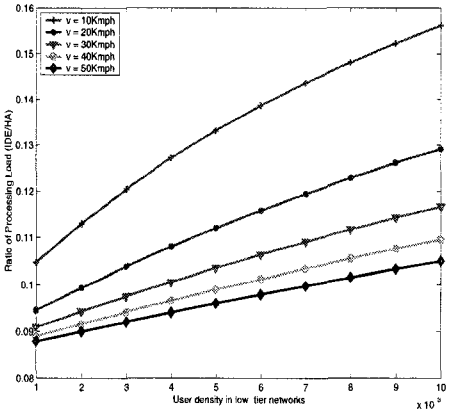

Figure 4.17 Ratio of processing load vs. user density.

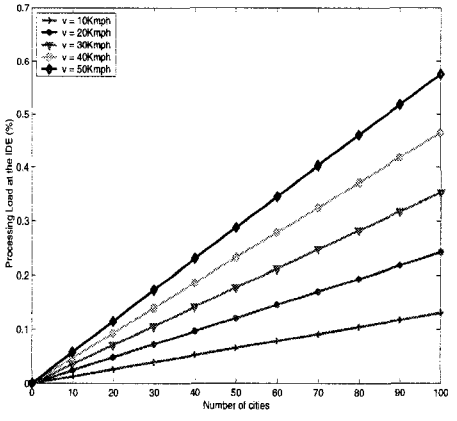

Figure 4.18 Processing load at the IDE vs. number of cities.

4.7 Conclusion

Mobility management and systems interworking are crucial in NGWN/4G. Several IPv6-based protocols have been proposed for mobility management at the IP

layer. However, they cannot guarantee seamless roaming and services continuity for real-time applications. On the other hand, interworking architectures available in the literature fail to fulfill all requirements for delay- and loss-sensitive applications.

In order to enable a better network performance in heterogeneous IP-based wireless and mobile environments, this paper proposes a novel interworking architecture, called *Integrated InterSystem Architecture* (IISA), and a handoff management protocol, called *Handoff Protocol for Integrated Networks* (HPIN). The proposed interworking architecture, IISA, is based on an adaptive loose coupling approach and introduces a third-party entity, called the *Interworking Decision Engine* (IDE), in order to guarantee the seamless roaming and services continuity required in NGWN/4G. Moreover, IISA has several advantages such as scalability, easy deployment and it supports roaming between various heterogeneous wireless networks.

The HPIN is a one-suite protocol that carries out access networks discovery, fast handoff and localized mobility management. HPIN reduces service disruption during a handoff by anticipating the handoff and allowing the selection of the best available network. The performance analysis demonstrates significant gains for quality of service (QoS) defined in terms of signaling overhead, handoff latency, packet loss and handoff blocking probability than current mobility management protocols. IISA and HPIN can guarantee seamless handoff, services continuity and QoS for an MN roaming across heterogeneous IP-based wireless environments. Furthermore, HPIN and IISA are simple enough, thus, their deployment will not require strong effort and extensive costs. Future work is to validate numerical results by using intensive simulations and prototype.

CHAPITRE 5

ADAPTIVE HANDOFF SCHEME FOR HETEROGENEOUS IP WIRELESS NETWORKS

Christian Makaya and Samuel Pierre
Mobile Computing and Networking Research Laboratory (LARIM)
Department of Computer Engineering, École Polytechnique de Montréal
P.O. Box 6079, Station Centre-ville, Montréal, Québec, H3C 3A7, Canada
Email : {christian.makaya, samuel.pierre}@polymtl.ca

Abstract

Recent technological advances allow mobile devices to be equipped with multiple wireless interfaces. Moreover, the coexistence of diverse but complementary architectures and wireless access technologies consist of a major trend in 4G or next generation wireless networks (NGWN/4G). In this context, the selection of an appropriate interface to ensure that a mobile node (MN) remains connected to the network is a challenging issue for seamless roaming. Furthermore, mobility management as well as the integration and interworking of existing wireless systems are a complex task due to their specific characteristics. This paper proposes an efficient handoff protocol, called *Handoff Protocol for Integrated Networks* (HPIN), which alleviates services disruption during handoff in NGWN/4G. HPIN is based on a novel handoff decision function and carries out localized mobility, fast handoff and access networks discovery. Performance evaluation based on numerical results shows that the proposed scheme performs better than existing schemes.

Keywords : Mobility management, quality of service (QoS), next generation wi-

reless networks, vertical handoff, interworking architecture, seamless roaming.

5.1 Introduction

Next generation or 4G wireless networks (NGWN/4G) are expected to exhibit heterogeneity in terms of wireless access technologies, personalized and user-oriented services, application requirements, high usability and increased capacity. With NGWN/4G, users will have greater demands for seamless roaming across different wireless networks, support of various services (e.g., multimedia applications) and quality of service (QoS) guarantees. The advantages of 3G cellular networks, such as UMTS and CDMA2000, reside in of their global coverage while their weaknesses lie in their bandwidth capacity and operation costs. On the other hand, WLAN technology, such as IEEE 802.11, offers higher bandwidth coupled with low operation costs, although it covers a relatively short range. These existing wireless networks have been subject of extensive individual investigations. Moreover, technological advances in the evolution of portable devices made possible the support of different radio access technologies (RATs) under multi-homing concepts. This has raised much interest for the integration and interworking of 3G wireless networks and WLAN capable of providing integrated authentication, billing and global roaming. Users will have exactly one service subscription with one service provider in order to benefit connection anytime and anywhere, known as *always best connected* (Gustafsson & Jonsson, 2003).

The integration of these existing systems seems unavoidable due to the potential benefits of their complementarity and will be the basis of NGWN design rather than invest efforts into developing new radio interfaces and technologies Hui & Yeung (2003). An integrated and interworking architecture for NGWN should handle specific requirements and satisfies the following main features (Akyildiz *et al.* , 2005) :

economical, scalable, provision of seamless mobility and security. Conceptually, a typical NGWN framework can be viewed as many overlapping wireless access domains, as shown in Fig. 5.1. Heterogeneity in terms of RATs and network protocols in NGWN requires a common interconnection element. Since the Internet Protocol (IP) technology enables the support of applications in a cost-effective and scalable way, it is expected to become the core backbone network of NGWN (Akyildiz *et al.* , 2005). Thus, current trends in communication networks evolution are directed towards an all-IP principles in order to hide heterogeneity and to achieve convergence of various networks. For example, third generation wireless initiatives, 3GPP and 3GPP2, adopted IPv6 as the sole IP version for the IP-based Multimedia Subsystem (IMS) (Chen & Zhang, 2004).

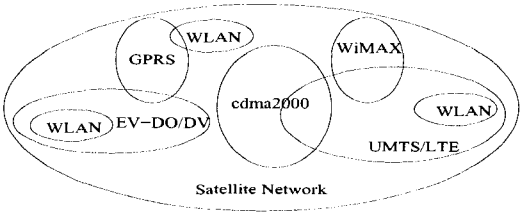

Figure 5.1 Overview of 4G/NGWN architecture.

Mobility management, with provision of seamless handoff and QoS guarantees to mobile nodes (MNs), is one of the key topics in NGWN/4G. It is crucial to provide seamless mobility and services continuity (i.e., minimal service disruption during roaming) support based on intelligent and efficient techniques. This means that seamless handoff schemes should have following features : minimum handoff latency, lower packet loss, limited handoff failure or blocking and lower signaling overhead. The handoff latency refers to the time interval during which an MN cannot send or receive any data traffic during handoffs. It is composed of L2 (link switching) and L3 (IP layer) handoff latencies. The overall handoff latency may be sufficiently long and leads to packet loss, which is inappropriate for real-time

applications such as voice over IP (VoIP). The signaling traffic overhead is defined as the total number of control packets (for registration, binding update and binding refresh procedures) exchanged between an MN and a mobility agent (e.g., home agent).

The handoff process in NGWN is composed of three phases : network discovery, handoff decision and handoff execution. The most simple way for a multiple interfaces MN to discover reachable wireless networks is to keep all air-interfaces on at all times. However, keeping an air-interface active all the time consumes battery power and bandwidth even when the device unit is not sending or receiving any packets. The handoff decision refers to the process of selecting the right moment when to perform the handoff. It is thus critical to avoid keeping idle air-interfaces perpetually on. Moreover, in order to avoid the ping-pong effect, an MN must observe if the new network is consistently better than the current one before performing a handoff.

In homogeneous networks, the handoff decision is typically driven by metrics which are strictly related to the received signal strength (RSS) level and resources availability. However, in NGWN, the RSS from different networks do not have the same meaning since each network is composed of its specific characteristics and there is no common pilot. Then, RSS comparisons are insufficient for handoff decision and may be inefficient or impractical. A more complex decision criterion that combines a large number of parameters or factors such as monetary cost, bandwidth, power consumption and user profile is necessary.

This paper proposes a novel mobility management scheme, called *Handoff Protocol for Integrated Networks* (HPIN), that enables QoS guarantee for real-time applications in heterogeneous IPv6-based wireless environments. HPIN is a one-suite protocol that performs network selection based on our proposed handoff score

function approach. Moreover, HPIN performs fast handoff, localized mobility management, context transfer and access network discovery. The aim of HPIN is to allow seamless roaming and services continuity across various access networks. The remainder of this paper is organized as follows. The following section offers an overview of basic concepts and work related to interworking and mobility management in heterogeneous IP wireless networks. After that, an interworking architecture, called *Integrated InterSystem Architecture* (IISA) is presented along with HPIN. Subsequently, an analytical framework to derive a signaling traffic cost, handoff latency and total packet loss is described. The performance analysis based on this analytical framework is carried out before concluding remarks.

5.2 Background and Related Work

Mobility management enables systems to locate roaming terminals in order to deliver data packets (i.e., *location management*) and maintain connections with them when moving into new subnet (i.e., *handoff management*). With the coexistence of various wireless access technologies, two kinds of handoffs are possible in NGWN : *horizontal* and *vertical handoffs*. Horizontal or intrasystem handoff occurs when an MN moves between the access points (APs) or base stations (BSs) of a same network technology. When AP/BSs belong to different networks (e.g., IEEE 802.11 and UMTS), such movement refers to vertical or intersystem handoff. NGWN characteristics make the implementation of vertical handoff more challenging than horizontal handoff. In fact, maintaining an uninterrupted session while the physical interface changes is very complex.

An evident way to achieve roaming among networks of different service providers or network operators consists of using Service Level Agreements (SLAs). However, due to several reasons, this approach is not always feasible. In fact, the increasing

number of wireless networks make it impractical for network operators to have direct SLAs with every single operator. Moreover, the SLAs can only provide static information. Furthermore, the network operators are reticent to the idea of opening their databases to others. With characteristics of mobility in NGWN/4G the user profile seems to be important when performing handoff decision. More complex metrics combining a large number of parameters such as monetary cost, bandwidth, power consumption, service types, network conditions and user preferences should be defined for handoff decision in NGWN (McNair & Zhu, 2004). Designing handoff decision function to evaluate these various metrics simultaneously is crucial in NGWN and remains a challenging research issue.

Various schemes for horizontal handoff have been proposed in the literature (Akyildiz *et al.* , 1999). Recently, research on vertical handoff in NGWN/4G attracted more attention and some works have been presented in the literature with their strength and weaknesses (Zhu & McNair, 2006). Several of these related papers use a handoff decision based on RSS and bandwidth. On the other hand, other proposals focus on the design of an architecture for heterogeneous networks such as the IPv6-based mobility schemes proposed by the Internet Engineering Task Force (IETF). Mobile IPv6 (MIPv6) (Johnson *et al.* , 2004) is proposed for mobility management at the IP layer and allows MNs to remain reachable in spite of their movements within IP-based mobile environments. Each MN is always identified by its home address, regardless of its current point of attachment to the network. While away from its home network, an MN is also associated with a care-of address (CoA), which provides information about its current location. However, MIPv6 has some well-known drawbacks such as signaling traffic overhead, high packet loss and handoff latency, thereby causing a user-perceptible deterioration of real-time traffic (Pérez-Costa *et al.* , 2003; Gwon *et al.* , 2004).

These weaknesses led to the investigation of other solutions to enhance MIPv6.

Two main MIPv6 extensions proposed by the IETF are the Hierarchical MIPv6 (HMIPv6) (Soliman *et al.* , 2005) and the Fast Handovers for MIPv6 (FMIPv6) (Koodli, 2005). These protocols tackle intra-domain or micro-mobility while MIPv6 is used for inter-domain or macro-mobility. HMIPv6 handles handoff locally through a special node called Mobility Anchor Point (MAP). The MAP acts as a local home agent (HA) in the network visited by an MN, limits the amount of MIPv6 signaling outside its domain and reduces the location update delay. However, HMIPv6 cannot meet the requirements of delay-sensitive traffic such as voice over IP (VoIP), due to packet loss and handoff latency (Pérez-Costa *et al.* , 2003; Gwon *et al.* , 2004). FMIPv6 was proposed to reduce handoff latency and minimize services disruption during handoff pertaining to MIPv6 operations such as movement detection, binding update and addresses configuration. In other words, FMIPv6 allows an MN to receive data before the binding is done at the HA and correspondent nodes (CNs). The link layer information (*L2 trigger*) is used either to predict or respond rapidly to handoff events.

Although FMIPv6 paves the way for improving MIPv6 performance in terms of handoff latency, it does not efficiently reduce signaling overhead (due to new messages introduced and exchanged for handoff anticipation) nor does it prevent packet loss (due to buffer space requirement). This may lead to unacceptable service disruption for real-time applications. Combining HMIPv6 and FMIPv6 motivates the design of Fast Handover for HMIPv6 (F-HMIPv6) (Jung *et al.* , 2005a) to increase network bandwidth usage efficiency. However, F-HMIPv6 may inherit drawbacks from both FMIPv6 and HMIPv6, for example synchronization and signaling overhead issues. In fact, in F-HMIPv6, when an MN performs a handoff immediately after sending a fast binding update (FBU) message to the MAP, all packets transferred to the previous on-link care-of address (PLCoA) during the period that the FBU needs to reach to the MAP, are lost (Pérez-Costa *et al.* , 2003). Moreover,

F-HMIPv6 provides fast handoff and localized mobility management although, it does not provide context transfer, access router and network discovery in the same way as for FMIPv6 and HMIPv6.

An architecture for next generation all-IP-based wireless systems is proposed in (Akyildiz *et al.* , 2005), called Architecture for Ubiquitous Mobile Communications (AMC). Two new entities, the Network Interworking Agent (NIA) and the Interworking Gateway (IG), are introduced in order to allow the integration of several wireless networks while supporting MN roaming. Moreover, an intersystem handoff protocol at the IP layer is designed for mobility management in this new architecture. However, the AMC architecture provides no appropriate handoff decision mechanism to take heterogeneity into account. The deployment of IG entity in all networks may require excessive economical costs and require changes in individual networks. Furthermore, the AMC architecture is based only on SLAs which can provide only static information. On the other hand, AMC may inherit certain drawbacks of loose coupling. The handoff decision is based on RSS criterion, which is inappropriate for NGWN as stated above. Also, air-interfaces always on approach is used in AMC architecture. The QoS provision and guarantees are not taken into account in AMC.

Other works have been presented in Akyildiz *et al.* (2004); Shenoy (2005) and Assouma *et al.* (2006) for intersystem mobility management and interworking of heterogeneous 3G cellular wireless networks, yet not for IP-based heterogeneous wireless networks. Often, proposed integration schemes are based on the deployment of a gateway, which solves interworking issues between each pair of networks. Adding a gateway at the boundaries of both systems would increase deployment costs. Moreover, these studies seem to integrate only cellular networks. To reduce energy consumption of MNs without degrading throughput, an approach called WISE (Wise Interface SElection) (Minji *et al.* , 2004) for 3G/WLAN vertical han-

doff has been proposed. With WISE, the handoff decision is performed according to the network load and the energy consumption of the air-interfaces. However, requirements such as services and applications security are not considered. In Buddhikot *et al.* (2003), an integrated architecture and interface selection schemes are proposed based on signal strength and radio interfaces priorities. As aforementioned, these parameters are not appropriate for handoff decision in NGWN. Moreover, an MN must passively evaluate handoff conditions, even when the application is running well in the current network. This introduces unnecessary power consumption and usage of network resources.

The IETF proposed a policy-based architecture in order to implement a set of rules to manage and control access to network resources which is particularly useful for QoS management (Yavatkar *et al.* , 2000). Two main logical entities for policy control-based architecture are the Policy Decision Point (PDP) and the Policy Enforcement Point (PEP). To enable judicious choice for vertical handoff, several papers have proposed a cost function to measure the network quality. A policy-enabled handoff decision algorithm proposed in Wang *et al.* (1999) is based on a cost function approach that considers several factors (e.g., bandwidth, power consumption and monetary cost). This cost function is very simple and cannot handle more sophisticated scenarios.

Moreover, the cost function evaluation could require high processing time and power. A vertical handoff decision algorithm has been proposed in McNair & Zhu (2004) and metrics that characterize NGWN have been identified. However, the proposed cost function could lead to singularity problems if connections are free of charge. Furthermore, handoff instability problem and mobility management at the IP layer are ignored. The factors considered in the above cited papers are insufficient. In fact, information about authentication types, access network types and the support of roaming partners are not taken into account. Moreover, these

studies do not provide a viable architecture framework for selection mechanisms, nor business models for prospective deployment.

5.3 Interworking Architecture for NGWN

As stated in 3GPP (2003), no use cases have been identified for the access to 3G wireless system circuit-switched based services scenario. Thus, for further development, it is not considered worthwhile. Hence, we focus on two main scenarios : service continuity and seamless services provision. Based on 3GPP/3GPP2-WLAN interworking models, an interworking architecture, called *Integrated InterSystem Architecture* (IISA) is proposed in Makaya & Pierre (2007b) and is shown in Fig. 5.2. For the sake of simplicity, only UMTS, CDMA2000 and WLAN networks are illustrated. Although IISA is designed to integrate any number of radio access technologies (RATs) and mobile devices may be equipped with any number of interfaces. Instead of developing new infrastructures, IISA extends existing infrastructures to tackle integration and interworking issues and provide mobile users with ubiquity or always best connected (Gustafsson & Jonsson, 2003).

The serving GPRS (general packet radio service) support node (SGSN) and packet control function (PCF) are enhanced with the AR functionalities and called *Access Edge Node* (AEN). Similarly, the gateway GPRS support node (GGSN) and packet data serving node (PDSN) are extended with MAP or HA functionalities (to enable message format conversion, QoS requirement mapping, etc.) and is called *Border Edge Node* (BEN). The WLAN Interworking Gateway (WIG) acts as a route policy element, ensuring message format conversion. Extended functionalities can be integrated into the existing networks entities or implemented separately. We advocate the first choice as it is more easily managed and implemented. Mapping between the home location register or the home subscriber server (HLR/HSS) in 3G

wireless networks and AAA server in WLAN is required to execute authentication and billing when user roams across both technologies.

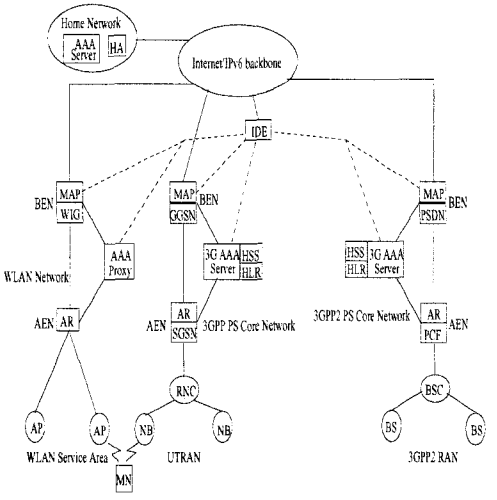

Figure 5.2 Integrated InterSystem Architecture (IISA).

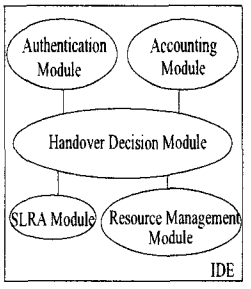

Figure 5.3 Interworking Decision Engine (IDE).

A novel entity, *Interworking Decision Engine* (IDE) shown in Fig. 5.3, is introduced to enable the interworking and handoff between various networks by reducing signaling traffic, services disruption during handoff and it also handles authentication, authorization and accounting (AAA) as well as mobility management. The usage of the IDE could be seen as a value-added service that network operators offer to their subscribers to allow roaming to other networks. To avoid additional signaling overhead due to the execution of AAA procedure every time an MN performs a handoff and requests registration, a token-based approach is proposed. This token is obtained from the IDE after the MN first successful registration in the foreign network. The token includes security association parameters to setup a secure tunnel between an MN and AR/AENs. IISA allows the separation of control and transport plane. In fact, data packet traffic bypasses the IDE. In other words, the IDE is in a control plane while the MAP/BEN handles the actual data traffic,

thus it is in the transport plane.

5.4 Proposed Handoff Protocol

With the coexistence of various access networks, the selection of subsystem that allows better service provision to subscriber is crucial and depends on several factors, for example defined in user's profile and preferences. Then, a more complex handoff decision criterion combining a large number of parameters or factors such as monetary cost, bandwidth, priority, power consumption, service types, system performance, user preferences, MN moving speed, security, resource availability, network accessibility and MN conditions, must be defined in NGWN (McNair & Zhu, 2004). The design of handoff decision function which evaluates these various factors simultaneously is crucial in NGWN and remains a research challenge. A hybrid vertical handoff decision function in heterogeneous wireless networks is proposed in this paper to provide satisfactory overall performance, based on the aforementioned criterion.

5.4.1 Handoff Score Function

In NGWN, the selection of the best access network is important during connection or handoff request. Handoff triggering is performed either in the access network by an MN, AP/AR, while the whole handoff process may require several entities located either in the home and foreign network, particularly for vertical handoffs. In fact, handoff triggered by an MN and/or AP/AR could be conducted only with locally available information such as link quality, signal strength, AR's capabilities and subnet load and can lead to inefficient system performance. Other information recorded in the home and foreign networks such as operator policies, access net-

work's load and user preferences, can be relevant for network selection and handoff decision. This paper proposes a novel handoff decision function for this purpose.

The usage of network n at a certain time is associated to a score which is a function of several of the aforementioned parameters. Those parameters can co-relate, interact and may have conflicting objectives. The presence of such conflicts makes it difficult to find an effective solution that optimizes all criteria simultaneously. In this case, network selection issue can be formulated as a multicriteria optimization problem. Several approaches are available in the literature in order to solve multicriteria optimization problem. Amongst them, we use a weighted sum approach introduced in Yager (1988).

For a given user u, the score function (f_u^n) is evaluated for each network n that can provide user services. In other words, the score function quantifies the QoS provided by a wireless network to handle running application on an MN device. The target network that results in the least highest computed score function value among all candidates is the network that would provide significant benefits (i.e., QoS level) to the user. More specifically, let n_c be the current serving network, $\mathcal{N}$ denotes the set of neighbor networks of n_c and $\mathcal{F}_u$ represents a set defined by : $\mathcal{F}_u = \{f_u^n : f_u^n > f_u^{n_c},\ \forall n \in \mathcal{N}\}$. The optimal target network, n^*, for a mobile user u is obtained as follows :

$$f_u^{n^*} = \inf(\mathcal{F}_u) \tag{5.1}$$

where f_u^n is expressed by :

$$f_u^n = \sum_s p_{u,s}^n f_{u,s}^n \qquad \forall\ n,\ u. \tag{5.2}$$

The score function of unreachable network always equals zero. $p_{u,s}^n$ is the priority of service/session s for network n based on user u profile, i.e., the probability that

an MN prefers network n for a connection of service s and $f^n_{u,s}$ is the per-service score function for network n. In other words, it represents a QoS factor and is computed as follows :

$$f^n_{u,s} = \sum_i w^n_{s,i} f^{n,u}_{s,i} \qquad \forall s,\ n,\ u \tag{5.3}$$

where $f^{n,u}_{s,i}$ is the normalized QoS function/factor provided by network n for parameter i to carry out service s and $w^n_{s,i}$ stands for a weight indicating the impact of the QoS parameters on either user or network and sum to one, i.e., $\sum_i w^n_{s,i} = 1$ and $w^n_{s,i} \in (0, 1]$. The assignment of weights $w^n_{s,i}$ plays a key role in the network selection. Hence, it is assumed that the assignment of these weights is based on the subscribers' home network policy or users profile. The target network is the network which provides just enough consistently higher QoS level than current network. Due to dynamic network conditions of wireless environments, the score function of target and current wireless networks may vary considerably. Then, a dwell timer or stability period should be adjusted according to the current measurements of the handoff score function.

In order to reflect the inability of candidate networks to guarantee the desired QoS requirements and to speed up score function evaluation, several constraints are considered depending on each factor such as the MN speed, bandwidth threshold, ARs load and delay. Hence, it is necessary to define maximal and minimal requirements for each parameter to enable the application provision. Then, if an available network cannot guarantee a minimum requested QoS (e.g., delay for real-time applications or bandwidth), it should be immediately discarded as a candidate network when there are several networks available. Otherwise, the network which allows best effort as QoS level is selected. The processing time and power are then reduced during the computation of the score function. The normalized QoS function

$f_{s,i}^{n,u}$ is given by :

$$f_{s,i}^{n,u} = \begin{cases} 0 & \text{if } Q_{s,i}^{n,u} \leq L_{s,i}^{n} \\ \left(\dfrac{Q_{s,i}^{n,u} - L_{s,i}^{n}}{U_{s,i}^{n} - L_{s,i}^{n}} \right)^{\alpha_i} & \text{if } L_{s,i}^{n} < Q_{s,i}^{n,u} < U_{s,i}^{n} \\ 1 & \text{if } Q_{s,i}^{n,u} \geq U_{s,i}^{n} \end{cases} \tag{5.4}$$

where $Q_{s,i}^{n,u}$ is the real value of parameter i in wireless network n associated to application s, measured by the MN or announced by a mobility agent.

$L_{s,i}^{n}$ and $U_{s,i}^{n}$ respectively express the minimal and maximal requirement of parameter i associated with wireless networks n for application s. These boundaries make it possible to check if the serving network satisfies the application's requirements. α_i is a constant that can take different values in order to specify different normalized QoS functions for each parameter i. The values of α_i greater than 1 result in a slow increase from the unacceptable required boundaries and fast near the maximal required boundary. If α_i equals 1, the normalized QoS function is strictly proportional between the required boundaries. Values of α_i lower than 1 result in a fast increase from the unacceptable required boundaries and slow near the maximal required boundary. Normalization is needed to ensure that the sum of the values, measured with different, units is meaningful.

5.4.2 Handoff Decision Algorithm

The proposed score function for handoff decision may be computed at MN side or at the IDE. In fact, assuming that mobile devices will become increasingly powerful, intelligent and sensitive to link layer changes, we adopt a network-assisted and mobile-controlled handoff strategy. The proposed handoff scheme combines mobile-monitored and network-probed information to provide reliable handoff management. Prior to handoff, an MN can obtain information from wireless network

candidates to which it is likely to handoff, and use such information to optimize handoff performance. On the other hand, if mobile device capabilities are limited, the handoff score function is computed at the IDE. In this case, handoff strategy turns to mobile-assisted and network-controlled.

The sequential execution of the system discovery and handoff decision steps for vertical handoff may be inappropriate. In fact, if the ongoing session of an MN runs with satisfactory QoS level with the current network, there is no need to discover another better network. Performing unnecessary handoff operations will waste network resources and energy of MN battery. We propose a handoff scheme based on a cross system discovery and handoff decision steps. In this proposed scheme, the current network conditions are checked first after an impending handoff event is generated if they can satisfy the ongoing session requirements. If they do, there is no need to perform the system discovery process, which will be launched only if the network conditions cannot satisfy an MN session requirements. Otherwise, if possible an MN tries to perform horizontal handoff. However, if no other AR/AEN of current network or technology exists, a vertical handoff is needed for this MN and it tries to find another more suitable network.

To avoid all air-interfaces always on approach for system discovery, we propose an adaptive scheme. An MN requests neighbors networks information from its serving network to the IDE. Through information reported periodically to the IDE, it maintains a global view of the connection state of roaming MNs and access networks conditions in its coverage area. The IDE replies by sending information about neighbor networks, if any, to the MN through its point of attachment. Then, the MN will compute the handoff score function for each reachable network using the information received in order to determine candidate networks. If candidate networks are available, the MN sets up a waiting timer to assess the stability of these candidate networks. If the QoS level remains better until the waiting timer

expires, the MN selects the candidate network which offers a QoS level that is slightly better than the current serving network and can start the handoff execution step. Otherwise, it will remain in the current network as the target network is unstable and cannot maintain better QoS level during the waiting time.

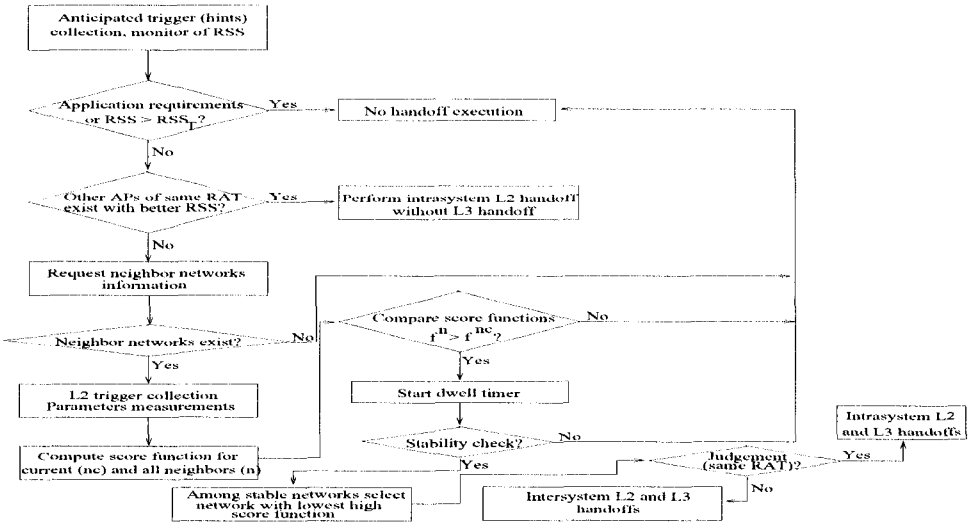

Figure 5.4 Flow chart of handoff decision algorithm.

The choice of candidate network that offers a slightly improved QoS level overcomes inefficient usage of network resources. The handoff decision algorithm flow chart is illustrated in Fig. 5.4, where RSS_T is a predefined or adaptive received signal strength threshold value. Although it is possible to select two different networks to access two ongoing services (in case of multiservice/session) simultaneously, in practice such a choice can create several problems. For example, authentication with two networks simultaneously, turning on two radio-interfaces at the same time and routing of service data appropriately within the device and the network represent some of these challenges. It is thus recommended to avoid such network selection approach.

5.4.3 Operation Mode of HPIN

Usually, handoff schemes assume that an MN monitors periodically neighbor AR/AENs signal strength by keeping all of its interfaces always on. However, keeping on all interfaces continuously drains the MN battery energy. This problem becomes worst when the number of RATs supported by the mobile devices increases and are available in an MN's moving area. To achieve seamless mobility across various access technologies and networks, an MN needs the information about the wireless network to which it could attach. It is also necessary to transfer information (context transfer) related to the MN from the current AR/AEN to the next one. The proposed *Handoff Protocol for Integrated Networks* (HPIN) implements access router and network discovery based on message exchanged between the IDE and mobility agents, minimizes the usage of limited wireless resources, provides fast mobility and secure transfer. The key objective of HPIN is to reduce services disruption and to avoid the re-initiation of signaling to/from an MN during handoff from the beginning.

5.4.3.1 Overview of HPIN

The *Router Information eXchange* (RIX Request/Reply) messages are used to allow the MAP/BEN and the IDE to maintain a global view of their coverage area. The RIX messages exchange is quite similar to the routing information protocol (RIP) (Hedrick, 1988) which allow neighboring routers to exchange their routing tables with one another. The information about the global view may be defined in terms of system parameters, subnet load, QoS status and information (e.g., supported data rate, video coding rate, maximum delay), bandwidth availability, and MN's signal strength. Four main messages are introduced for handoff management :

- *Handoff Preparation Request* (HPReq) message sent from an MN to the

MAP/BEN for a handoff request. It contains information about user preferences/profiles, applications required QoS capabilities, L2 information of AR/AENs, IP address of MNs, signal strength of the MN, AR/AEN's ID for an MN location tracking.

- *Handoff Preparation Reply* (HPRep) message sent by the MAP/BEN to an MN. It contains network prefixes, the list of candidates AEN (CAR/AEN), their capabilities and the QoS status.
- *Handoff Preparation Notification* (HPN) message sent by an MN to the MAP/BEN to notify the possibility of an impending handoff. It contains the information about the selected new AR/AEN where an MN will handoff. The HPN includes the request to verify the pre-configured new on-link care-of address (NLCoA) and establish a bi-directional tunnel between the MAP/BEN and the NAR/AEN in order to prevent routing failures during handoff.
- *New Link Attachment (NLA)* message sent by an MN to the NAR/AEN to announce its presence on the new link and confirm usage of the NLCoA.

Moreover, there are also two optional messages :

- *Handoff Preparation Acknowledgment* (HPAck) message which contains information about the current capabilities that an AR/AEN can support;
- *New Link Attachment Acknowledgment* (NLAck) message to notify an MN to use another NLCoA rather than its prospective NLCoA, in case of address collision. This message is also sent to the current MAP/BEN to allow it to bind previous on-link care-of address (PLCoA) and to validate the NLCoA.

The MN decides whether to send a request message (HPReq) for handoff preparation depends on the generation of anticipated triggers (AT). The high bit error rate, link going down, weak signal strength, security risks, monetary cost and geographical location can be used as anticipated triggers. To allow seamless service

continuity, requirements specified in the request message need to be set consistently with the QoS negotiated in the previous subnet. QoS consistency remains a very challenging issue and is crucial for real-time applications. This consistency is handled by the IDE, which allow QoS mapping between various networks. Mapping is needed to translate the QoS guarantees and specifications provided for a session across heterogeneous networks. The QoS mapping performed by the IDE is for example the requirements about resource reservation and delay threshold according to SLAs.

5.4.3.2 Roaming Scenarios

Upon receipt of HPReq message, the MAP/BEN checks its local CAR/AEN table to retrieve information about their capabilities. The MAP/BEN performs a pre-filtering process, based on the requirements specified in the HPReq and the available network conditions to obtain the CAR/AEN list. Note that, if the context information of this MN is not available at the candidate MAP/BEN, the latter sends Context Transfer Request (CTReq) message (Loughney *et al.* , 2005) to the IDE in order to get session management parameters of the MN for establishment of traffic bearers on the new path. In response to a CTReq message, the IDE transmits a Context Transfer Data (CTD) message that includes the MN's feature contexts.

When the new MAP/BEN receives a CTD message, it installs the contexts as received from the IDE. The MAP/BEN responds to an MN by sending a HPRep message. If the MAP/BEN lacks information about the user profile, it requests this information to the IDE rather than to the MN's home network, which may be away of the current MAP/BEN domain. When the MN receives a HPRep message, it performs stateless address configuration (Johnson *et al.* , 2004) to get new on-link CoAs (NLCoAs) and knows L2 technologies provided by CAR/AENs to which it is

likely to handoff. A primary NLCoA will be associated to the selected network. The MN will activate only the air-interface associated to the CAR/AEN list, rather than setting all air-interfaces always on. This selective air-interface activation enables a better trade-off between system discovery time, power consumption efficiency and allows shorter scanning delay.

Whenever an MN receives a L2 trigger, it initiates a target AR/AEN selection among CAR/AENs. This selection is based on the handoff score function, f_u^n, proposed above. After the target AR/AEN selection process, an MN notifies the MAP/BEN that it is moving into a new subnet by sending a HPN message to allow the MAP/BEN to establish a binding between PLCoA and NLCoA and to buffer all incoming and in-flight packets (this avoid synchronization issue identified in basic IP mobility schemes) having PLCoA as destination address. The MAP/BEN forwards a HPN message to the new access router (NAR/AEN) and includes the CTD message. Upon reception of acknowledgment from the NAR/AEN, the MAP/BEN starts tunneling any packets (buffered and incoming) addressed to PLCoA towards NLCoA. When the NAR/AEN receives the HPN, it performs the duplicate address detection (DAD) procedure on the NLCoA and application requirements validation. Then, the NAR/AEN responds to the HPN with a HPAck message. When the CAR/AEN processes the CTD message, it can optionally generate a CTD Reply (CTDR) message to report on the status of processing the received contexts and piggybacks this message in HPAck message.

After transmitting a HPN message, the MN performs a link layer switching and announces its presence on the new link by sending the NLA message to the NAR/AEN. Then, the NAR/AEN will start delivering the buffered packets to the MN. The packets forwarding procedure remains active until the binding update (BU) procedure is completed. Note that, if the HPN is not sent before the L2 handoff, the MN sends a HPN piggybacked in the NLA message (NLA[HPN]) over

the new link. This situation corresponds to the reactive mode of HPIN in contrary to its predictive mode, i.e., the HPN is sent through the previous link. When the NAR/AEN receives the NLA[HPN] message, it processes the NLA message part, extracts the HPN message part and forwards it to the serving MAP/BEN. At this time, the serving MAP/BEN starts buffering all incoming and in-flight packets having PLCoA as destination address and forwards them toward the NLCoA. If an address collision occurs when the NAR/AEN processes the NLA message, it changes the prospective NLCoA to a valid NLCoA and includes it in a HPN message before forwarding it to the MAP/BEN and simultaneously the NAR/AEN sends a NLAck to the MN. Fig. 5.5 illustrates the messages sequence during intra-BEN roaming while Fig. 5.6 represents the messages sequence flow during inter-BEN roaming.

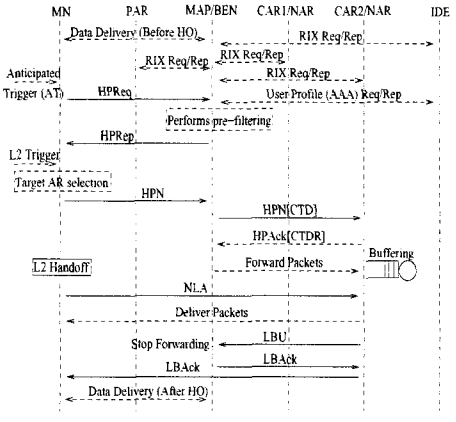

Figure 5.5 Signaling messages sequence for intra-BEN roaming.

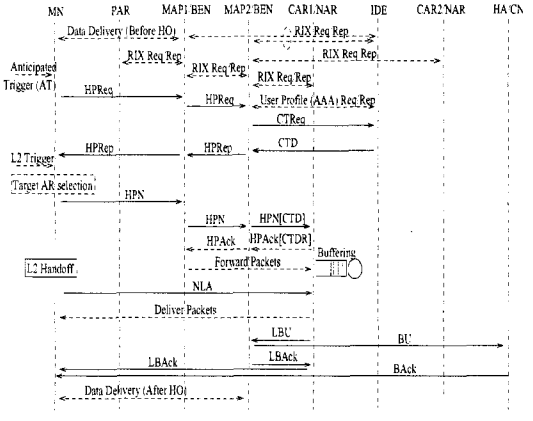

Figure 5.6 Signaling messages sequence for inter-BEN roaming.

The binding update (BU) procedure is performed by the NAR/AEN on the behalf of MNs. In fact, the AR/AEN acts as a proxy and copies a BU list of an MN in its cache and manages this list (e.g., lifetime entries) in the same way as the original is managed by the MN. The copy in the AR/AEN cache must be updated periodically according to the original BU list of the MN. As soon as an MN attached to the NAR/AEN, the copy of the BU list is used by the NAR/AEN to inform the

MAP/BEN about the NLCoA. When the lifetime of the BU list cached in the AR/AEN is about to expire, the AR/AEN can send a request for a BU list renewal to an MN. The BU list renewal is performed in the same way as basic BU refresh (Johnson *et al.* , 2004). The MN sends the BU list to NAR/AEN at the same time it attaches to the NAR/AEN. By piggybacking the BU list in a signaling message, separate out-of-band messages from MN to NAR/AEN are avoided, thus, reducing signaling traffic overhead.

5.5 Analytical Model for HPIN

In IP-based wireless networks, the QoS may be defined by packet loss, handoff latency and signaling overhead. Hence, analyzing of these metrics are most useful to evaluate the performance of mobility management protocols. The notation used in this paper is given in Table 5.1.

Tableau 5.1 Notation.

t_s	inter-session time between two consecutive sessions
t_c	subnet (AR/AEN's coverage area) residence time
t_d	MAP/BEN domain residence time
C^g	global binding update cost to HA/CNs
C^l	local binding update cost to MAP/BEN
N_{CN}	number of CNs with a binding cache entry for an MN
$d_{X,Y}$	average number of hops between nodes X and Y
$C_{X,Y}$	transmission cost of control packets between nodes X and Y
PC_X	processing cost for binding update at node X
P_s	probability of anticipated handoff signaling success
S_s	signaling cost for a successfully anticipated handover
S_f	signaling cost if no real L3 handoff occurs
S_r	signaling cost for reactive mode

5.5.1 User Mobility and Traffic Models

User mobility and traffic models are crucial for efficient system design and performance evaluation. Usually, an MN mobility is modeled by the cell residence time and various types of random variable are used for this purpose (Fang, 2003). Two commonly used mobility models in wireless networks are random-walk and fluid-flow models (Wang & Akyildiz, 2000). We consider the random walk, i.e., an MN moves at constant speed v with uniformly distributed angular directions belonging to $[0, 2\pi]$ as mobility model. Let $d_{i,u}$ be the distance between AP/BS i and mobile user u. We assume that the path loss or link gain is given by $h_{i,u} = 10^{\frac{s_{i,u}}{10}} d_{i,u}^{-\nu}$, where ν is the path loss exponent, $s_{i,u}$ is the log-normal shadowing with zero mean and standard deviation σ_s.

The exponential distribution provides an acceptable tradeoff between complexity and accuracy. Thus, most cost analyses adopt exponential assumption (Fang, 2003). We consider a traffic model with two levels, session and packets. The session duration follows an exponential distribution with inter-session rate λ_s while packet generation follows a Poisson process. Let μ_c and μ_d be the border crossing rate for an MN out of a subnet (i.e., AR/AEN domain) and a MAP/BEN domain, respectively. When an MN crosses a MAP/BEN domain border, it also crosses an AR/AEN border. Then, let μ_l be the border crossing rate for which an MN still stays in the same MAP/BEN domain, $\mu_l = \mu_c - \mu_d$.

If we assume that all subnets have circular shape and form together a contiguous area and that each MAP/BEN domain is composed of M equally subnets, we obtain : $\mu_d = \frac{\mu_c}{\sqrt{M}}$. The roaming probability depends on an MN's movement pattern in its original network but not in its destination network. Hence, the probabilities that there are at least one local binding update (P_c) and one global binding update

(P_d) between two consecutive sessions of an MN are :

$$P_c = Pr(t_s > t_c) = \frac{\mu_c}{\mu_c + \lambda_s} \text{ and } P_d = Pr(t_s > t_d) = \frac{\mu_d}{\mu_d + \lambda_s}. \quad (5.5)$$

The average number of location binding updates during an inter-session time corresponding to subnet crossings, $E(N_c)$, and MAP/BEN domain crossings, $E(N_d)$, are given by :

$$E(N_c) = \sum_{k=0}^{\infty} kP_c^k(1 - P_c) = \frac{\mu_c}{\lambda_s} \text{ and } E(N_d) = \sum_{m=0}^{\infty} mP_d^m(1 - P_d) = \frac{\mu_d}{\lambda_s}. \quad (5.6)$$

With the same time variables assumption, we can obtain the expression of $E(N_l)$, i.e., the average number of subnets that an MN crosses and still stay within a given MAP/BEN domain during an inter-session time interval, as follows : $E(N_l) = \mu_l/\lambda_s$.

5.5.2 Binding Update Signaling Cost

Performance analysis of wireless networks must consider the total signaling cost induced by a mobility management scheme. In NGWN, there are two kinds of location or binding update signaling. One occurs during an MN's subnet crossing while the other occurs when the binding is about to expire. Depending on the type of movement, two kinds of binding update can be performed : *global* and *local*. Global binding update occurs when an MN moves out of its MAP/BEN domain. In this case, the MN registers its new regional CoA (RCoA) to the HA and the CNs. On the other hand, if the MN changes its current address (LCoA) within a MAP/BEN domain, it only needs to register this new LCoA to the MAP/BEN. Hence, the average binding update signaling cost during inter-session time heavily

depends on the computation of numbers of binding updates and is given by :

$$C_{BU} = E(N_l)C^l + E(N_d)C^g = \frac{1}{SMR\sqrt{M}}\left[C^g + (\sqrt{M}-1)C^l\right] \tag{5.7}$$

where SMR is the session-to-mobility ratio and represents the relative ratio of session arrival rate over user mobility rate : $SMR = \lambda_s/\mu_c$.

Anticipated trigger and link layer information (L2 trigger) are used either to predict or rapidly respond to handoff events. Hence, HPIN signaling cost depends on the probability that the handoff anticipation is accurate. If there is no real handoff after the L2 trigger, all messages exchanged for handoff anticipation can be unnecessary. Thus, global and local binding update signaling cost for HPIN are expressed as follows :

$$\begin{aligned} C^g &= P_s S_s^g + (1-P_s)(S_f^g + S_r^g) + C_{ru} \\ C^l &= P_s S_s^l + (1-P_s)(S_f^l + S_r^l) + C_{mu} \end{aligned} \tag{5.8}$$

where C_{ru} and C_{mu} represent the binding update cost at HA/CNs and MAP/BEN, respectively. Their expressions are given in Table 5.2.

Tableau 5.2 Expression of partial signaling costs.

$C_{mu} =$	$2C_{AEN,BEN} + PC_{BEN}$
$C_{ru} =$	$2(C_{BEN,HA} + N_{CN}C_{BEN,CN}) + PC_{HA} + N_{CN}PC_{CN}$
$S_f^l =$	$C_{MN,BEN} + PC_{BEN} + PC_{AEN}$
$S_s^l =$	$C_{MN,BEN} + 2C_{BEN,AEN} + C_{MN,AEN} + PC_{BEN} + PC_{AEN}$
$S_r^l =$	$C_{MN,AEN} + C_{AEN,BEN} + PC_{BEN} + PC_{AEN}$
$S_f^g =$	$C_{MN,pBEN} + C_{pBEN,nBEN} + 2PC_{BEN} + PC_{AEN}$
$S_r^g =$	$C_{MN,AEN} + (C_{AEN,nBEN} + C_{nBEN,pBEN}) + 2PC_{BEN} + PC_{AEN}$
$S_s^g =$	$C_{MN,pBEN} + 2(C_{pBEN,nBEN} + C_{nBEN,AEN}) + C_{MN,AEN} + 2PC_{BEN} + PC_{AEN}$

5.5.3 Handoff Latency and Packet Loss

Since the number of packets lost is proportional to handoff latency, only the expression for handoff latency is derived in this section. The following parameters are defined to compute handoff latency and packet loss : t_{L2} the L2 handoff latency and $t_{X,Y}$ one-way transmission delay between nodes X and Y for a message of size s. If one of the endpoints is an MN, $t_{X,Y}$ is computed as follows :

$$t_{X,Y}(s) = \frac{1-q}{1+q}\left(\frac{s}{B_{wl}} + L_{wl}\right) + (d_{X,Y} - 1)\left(\frac{s}{B_w} + L_w + \varpi_q\right) \tag{5.9}$$

where q is the probability of wireless link failure and ϖ_q the average queueing delay for each router on the Internet (McNair *et al.* , 2001), B_{wl} (resp. B_w) represents the wireless (resp. wired) link bandwidth and L_{wl} (resp. L_w) denotes the wireless (resp. wired) link delay.

The HPIN handoff latency depends on the information available as well as the link where fast handoff messages are exchanged. The average handoff latency of HPIN for intra-MAP/BEN roaming is then given as follows :

$$D^l_{HPIN} = P_s O^l_{HPIN} + (1 - P_s) N^l_{HPIN} \tag{5.10}$$

where $O^l_{HPIN} = t_{L2} + 2t_{MN,AEN}$ is the handoff latency if the information about the NAR/AEN and impending handoff are available before the L2 handoff. Otherwise, this handoff latency is given by $N^l_{HPIN} = t_{L2} + 2t_{MN,AEN} + 2t_{AEN,BEN}$ associated to the HPIN reactive mode. For inter-MAP/BEN, N^l_{HPIN} becomes $N^g_{HPIN} = t_{L2} + 2t_{MN,AEN} + 2[t_{AEN,nBEN} + t_{nBEN,pBEN}]$ while $O^g_{HPIN} = O^l_{HPIN}$. In fact, handoff procedure only depends on intra-BEN communication delay, since the inter-BEN signaling is completed before the L2 handoff. The average handoff latency of HPIN for inter-BEN roaming is computed similarly as in (5.10).

5.6 Performance Evaluation

The performance analysis is conducted by examining several metrics such as throughput, handoff latency, packet loss and signaling traffic overhead. The parameter and default values used in the performance evaluation are listed in Table 5.3. An analytical framework to evaluate performance of IPv6-based handoff schemes proposed by the IETF (i.e., MIPv6, HMIPv6, FMIPv6 and F-HMIPv6) is presented in Makaya & Pierre (2007a). Such evaluation method is used to compare the performance of the IETF's protocols with HPIN. Traditional handoff protocols based on received signal strength (RSS) are compared with a handoff score function-based approach (SFA) used in HPIN. For the sake of simplicity, four parameters are used for network selection : power consumption (p), bandwidth (b), latency (l) and usage cost (c). Values used for those parameters and application requirements are given in Table 5.4.

Tableau 5.3 System parameters for performance evaluation.

Parameters	Symbols	Values
Wired link bandwidth	B_w	100 Mbps
WLAN bandwidth	B_{wl}	5.5 Mbps
UMTS bandwidth	B_{wl}	384 Kbps
Wired link delay	L_w	2 ms
Wireless link delay	L_{wl}	10 ms
Prediction probability	P_s	0.90
Wireless link failure probability	q	0.50
Control packet size	s_c	96 bytes
Data packet size	s_d	200 bytes
MN's average speed	v	5.6 Km/h
Time slot length	TS	5 s
Path loss exponent	ν	4
Shadowing standard deviation	σ_s	8 dB

The network topology considered for the analysis is illustrated in Fig. 5.7. It

Tableau 5.4 Network parameters and application requirement values.

	Network Parameters				Application Requirements					
	WLAN		UMTS		Idle		Voice		Data	
	$L^n_{s,i}$	$U^n_{s,i}$	$L^n_{s,i}$	$U^n_{s,i}$	$L^n_{s,i}$	$U^n_{s,i}$	$L^n_{s,i}$	$U^n_{s,i}$	$L^n_{s,i}$	$U^n_{s,i}$
Power (hour)	0	4	0	4	0	4	0	2	0	2
Latency (ms)	100	5	150	5	400	10	150	5	250	5
Bandwidth (Kbps)	250	800	20	200	3	20	9.6	64	50	500
Usage cost ($/min)	0	0.3	0.1	0.5	0	0	0.1	0.2	0.4	0.6

is assume that the distance between different domains is equal, i.e., $c = d = e = f = 10$ and set $a = 1$, $b = 2$, and $g = 4$. All links are supposed to be full-duplex in terms of capacity and delay. Parameter values used to compute signaling cost are defined as follows : $M = 2$, $\tau = 1$, $\kappa = 10$, $PC_{AEN} = 8$, $PC_{HA} = 24$, $PC_{CN} = 4$, $PC_{IDE} = 15$ and $PC_{BEN} = 12$. We assume that 3G/UMTS wireless

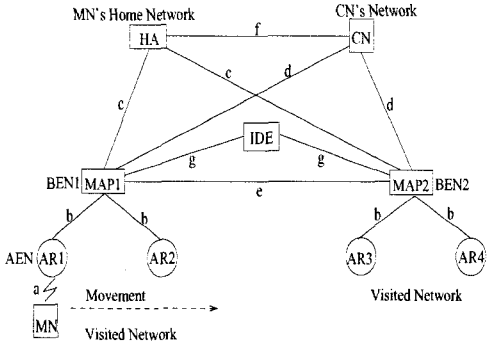

Figure 5.7 Network topology used for analysis.

networks overlap with WLAN (i.e., IEEE 802.11) networks and MNs give more weight to bandwidth and latency requirements, $w^n_{s,b} = w^n_{s,l} = 0.35$, followed by power consumption, $w^n_{s,p} = 0.20$ and less weight for usage cost $w^n_{s,c} = 0.10$ for all n and $\alpha_i = 0.3$ for all i. The performance analysis is conducted through MATLAB and OPNET softwares.

5.6.1 Throughput and Signaling Overhead

Fig. 5.8 shows the throughput (in packets/time slot) for both handoff decision schemes (RSS and SFA) when the average arrival rate of packets is 5 packets per second per user. A significant gain in throughput can be achieved with SFA/HPIN comparatively to the RSS scheme. The target MN is initially connected to UMTS, then, it moves towards the first WLAN, after it enters in overlapping area of all networks and moves into the second WLAN before returning to UMTS. When the target MN is located in the overlapping area, we can see how SFA/HPIN allows an increasing throughput compared to RSS scheme. In fact, with the RSS scheme, UMTS is chosen more often as a target network, since it provides highest signal strength and wide coverage area. This leads to negative side effects such as lower achievable data rate and imbalanced load. However, with the SFA/HPIN scheme, the subnet load can be efficiently distributed amongst all networks, leading to higher throughput guarantees. After the session switching from UMTS to WLAN, the throughput increases since the WLAN provides better network conditions and a higher packet rate. With varying packet arrival rate, Fig. 5.9 shows throughput ratio which refers to the ratio of the actual data rate over the requested rate. The SFA/HPIN scheme provides a better performance than the RSS, except when networks usage is low or congested.

To alleviate packet loss, fast handoff schemes should support packet buffering and forwarding during handoff execution. Fast handoff schemes (FMIPv6 and F-HMIPv6) require more buffer space than MIPv6 and HMIPv6 since they start packets buffering and forwarding early. HPIN requires less buffer space than F-HMIPv6 as illustrated in Fig. 5.10. In this analysis, the required buffer space for one MN during the handoff procedure is considered. The required buffer space increases according to the number of MN performing handoff and in proportion

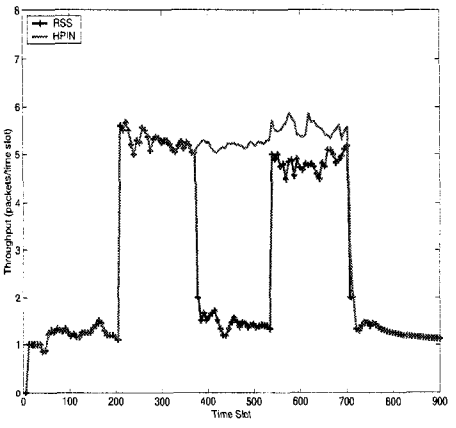

Figure 5.8 Target user throughput.

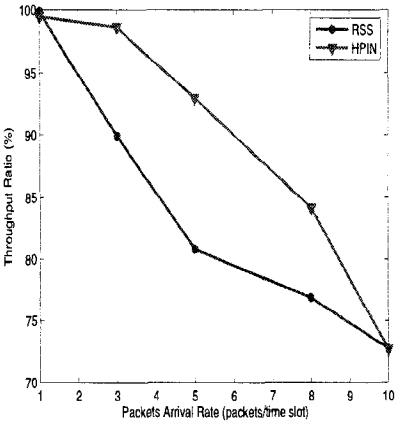

Figure 5.9 Throughput ratio comparison.

with the packet arrival rate. On the other hand, the buffering time may affect real-time applications. For example, if certain packets are stored in a buffer for a longer period of time than acceptable end-to-end delay, they may become useless. Hence, it is crucial to manage buffer efficiently in order to minimize overhead and provide better QoS to delay-sensitive applications.

Fig. 5.11 illustrates the signaling overhead cost during handoff as a function of the SMR. When the SMR is small, the mobility rate is larger than the session arrival rate. Then, an MN changes its point of attachment frequently due to its mobility, which results into several handoffs and increased signaling overhead. However, when the session arrival rate is superior to the mobility rate (i.e., $SMR > 1$), the binding update is less often performed and results into lower signaling overhead. FMIPv6 uses the wireless bandwidth more often than MIPv6 due to the additional messages it introduces for the handoff anticipation. For lower subnet residence time, the signaling overhead reduces considerably from FMIPv6 to HPIN. Furthermore, since the reactive mode of F-HMIPv6 correspond to HMIPv6, when an acknowledgment is not received by an MN through the previous link, the messages exchanged during

router discovery step becomes unnecessary. However, such messages exchange results in an increased in signaling overhead with F-HMIPv6 compared to HPIN. In fact, for F-HMIPv6, more messages are exchanged after the L2 trigger generation, which is not the case with HPIN. The RIX messages exchange introduces additional signaling similarly as with the routing information protocol (RIP). However, this signaling increment occurs only in the wired part of networks. Compared to the wireless part, the wired one has far much bandwidth and resources.

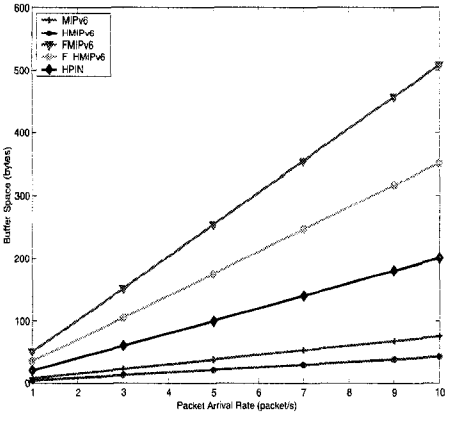

Figure 5.10 Impact of packet arrival rate on the required buffer space.

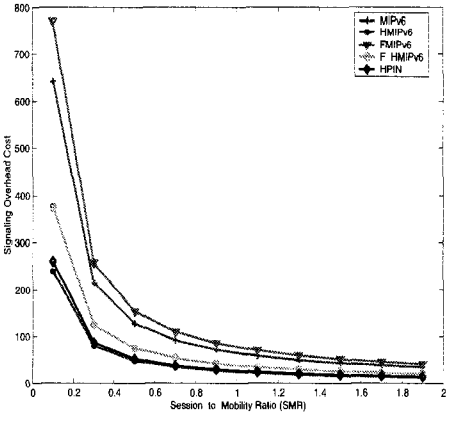

Figure 5.11 Binding update signaling traffic cost.

5.6.2 Handoff Latency and Packet Loss

According to Fig. 5.12, the handoff latency increases proportionally with the wireless link delay. We can observe that MIPv6 and HMIPv6 have the worst results among all protocols, followed by FMIPv6 and F-HMIPv6, while HPIN provides the lowest delay. In F-HMIPv6, the synchronization problem mentioned above is not solved and causes packet loss as well as increased data delay. This issue is solved in the HPIN, which allows a lower delay compared to F-HMIPv6. It is well known

that the maximal tolerable delay for interactive conversation is approximately 200 ms. Hence, HPIN can meet this requirement when the wireless link delay is set below to 50 ms. Since packet loss is proportional to handoff latency, similar results and behaviors are observed.

Fig. 5.13 shows the average packet loss versus the packet arrival rate. Packet loss is far lower for fast handoff schemes than for MIPv6 and HMIPv6. HPIN allows lower packet loss compared to other protocols. Due to the lack of any buffering and anticipated handoff mechanisms, all in-flight packets will be lost when the handoff is executed in MIPv6 and HMIPv6. However, in fast handoff schemes (FMIPv6, F-HMIPv6 and HPIN) packet loss begins from the moment the L2 handoff is detected until the buffering mechanism is initiated or if buffers overflow. This time interval is shorter for HPIN than for F-HMIPv6 due to its ability to solve the synchronization issue. Moreover, in HPIN, when the MN attaches to the new link, the re-directed packets are already waiting in the NAR/AEN.

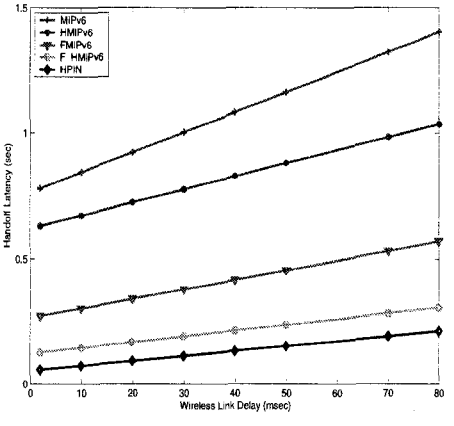

Figure 5.12 Impact of wireless link delay on handoff latency.

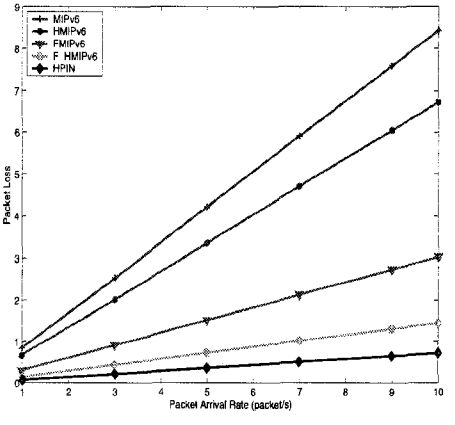

Figure 5.13 Impact of packet arrival rate on packet loss.

5.7 Conclusion

Mobility management and systems interworking are crucial in NGWN/4G. Several IPv6-based mobility management schemes have been proposed in the literature. However, they cannot guarantee seamless roaming and service continuity for real-time applications. On the other hand, interworking architectures described in the literature cannot fulfill all requirements for sensitive (e.g., delay and packet loss) applications. This paper proposes an efficient handoff management protocol, called *Handoff Protocol for Integrated Networks* (HPIN) to enable a better network performance in heterogeneous IPv6-based wireless environments.

HPIN is a one-suite protocol that performs access network discovery, context transfer, fast handoff and localized mobility mechanisms. An adaptive handoff decision scheme based on the score function derived by combining various criteria such as bandwidth, power consumption, latency and monetary cost is proposed. The HPIN provides guarantees for seamless roaming, services continuity and alleviates services disruption during handoff as required for NGWN/4G. Analyses of results from the performance evaluation indicate that HPIN improves performance in terms of throughput, handoff latency, packet loss and signaling overhead compared to other existing protocols, such as MIPv6, HMIPv6, FMIPv6 and F-HMIPv6. Plans for future work consist of validating numerical results using intensive simulation and testbed.

CHAPITRE 6

ENHANCED FAST HANDOFF SCHEME FOR HETEROGENEOUS WIRELESS NETWORKS

Christian Makaya and Samuel Pierre
Mobile Computing and Networking Research Laboratory (LARIM)
Department of Computer Engineering, École Polytechnique de Montréal
P.O. Box 6079, Station Centre-ville, Montreal, Quebec, H3C 3A7, Canada
Email : {christian.makaya, samuel.pierre}@polymtl.ca

Abstract

Mobility management, integration and interworking of existing wireless systems are important factors to obtain seamless roaming and services continuity in next generation or 4G wireless networks (NGWN/4G). Although, several IPv6-based mobility protocols as well as interworking architectures have been proposed, they cannot guarantee seamless roaming, especially for real-time applications. Moreover, mobility management protocols are designed for specific needs, for example, the purpose of IPv6-based mobility schemes consists of managing users roaming while ignoring access network discovery. This paper proposes an efficient handoff protocol, called *enhanced Handoff Protocol for Integrated Networks* (eHPIN), which carries out localized mobility management, fast handoff, and access network discovery. It alleviates services disruption during roaming in heterogeneous IP-based wireless environments. Performance evaluation shows that eHPIN provides better results in terms of signaling traffic overhead cost, handoff latency, packet delivery cost, handoff failure and packet loss compared to existing IPv6-based mobility schemes.

Keywords : Mobility management, IP mobility, quality of service, next-generation wireless networks, handoff, seamless roaming, service continuity.

6.1 Introduction

Next generation wireless networks (NGWN) or fourth generation wireless networks (4G) are expected to exhibit heterogeneity in terms of wireless access technologies, user-oriented services and greater capacities. Users will have increasing demands for seamless roaming across different wireless networks, support of various services (e.g., voice, video, data) and quality of service (QoS) guarantees. Hence, with this heterogeneity, users will be able to choose radio access technology (RAT) that offers higher quality, data speed and mobility which is best suited to the required multimedia applications. Moreover, technological advances in the evolution of portable devices make it possible to support different RATs. Heterogeneity in terms of RATs and network protocols in NGWN/4G requires common interconnection element. Since the Internet Protocol (IP) technology enables the support of applications in a cost-effective and scalable way, it is expected to become the core backbone of NGWN/4G (Akyildiz *et al.* , 2005). Thus, current trends in communication networks evolution are directed towards an all-IP principles in order to hide heterogeneities of lower-layers technologies from higher-layers and to achieve convergence of different networks.

Mobility management, with provision of seamless handoff and QoS guarantees, is one of the key topics in order to support global roaming of mobile nodes (MNs) in NGWN. Providing seamless mobility and service continuity (i.e., minimal service disruption during roaming) support based on intelligent and efficient techniques is crucial. This means that seamless handoff schemes should have following features : minimum handoff latency, low packet loss, low signaling overhead and limited han-

doff failure or blocking. Handoff latency represents the time interval during which an MN cannot send or receive any data traffic during handoffs. It is composed of L2 (link switching) and L3 (IP layer) handoff latencies. The overall handoff latency may be sufficiently long to cause packet loss, which is intolerable for real-time applications such as voice over IP (VoIP). Furthermore, subscribers are more sensitive to session/call blocking during handoff than to session blocking during session initiation. The handoff blocking probability refers to the likelihood that a session connection is prematurely terminated due to an unsuccessful handoff over a session lifetime. Hence, minimization of handoff blocking probability is crucial for mobility management schemes. The signaling traffic overhead is defined as the total number of control packets (for registration, binding update and binding refresh procedures) exchanged between an MN and a mobility agent (e.g., home agent).

Several IPv6-based mobility schemes such as Mobile IPv6 (MIPv6) (Johnson *et al.* , 2004), Hierarchical Mobile IPv6 (HMIPv6)(Soliman *et al.* , 2005) and Fast Handovers for Mobile IPv6 (FMIPv6) Koodli (2005), have been proposed by the Internet Engineering Task Force (IETF) to enable an MN to remain reachable when moving out of its home network. However, these protocols are hindered by several drawbacks such as signaling overhead, handoff latency and packet loss. To achieve seamless mobility across various access technologies and networks, an MN needs to have information regarding the wireless network to which it can attach. To enable this, Candidate Access Router Discovery (CARD) protocol (Leibsch *et al.* , 2005) was proposed by the IETF. When coupled with CARD protocol, traditional fast handoffs schemes may work inefficiently. Enhancing those protocols for efficient mobility management in heterogeneous NGWN networks is highly necessary.

This paper proposes a mobility management scheme, called *enhanced Handoff Protocol for Integrated Networks* (eHPIN), that enables seamless service continuity and QoS guarantees for real-time applications in heterogeneous IPv6-based wireless

environments. eHPIN performs access network discovery, localized mobility and fast handoff management. In other words, eHPIN aims to provide efficient access network discovery and roaming support in order to alleviate services disruption during handoff in NGWN/4G. The remainder of this paper is organized as follows. In Section 6.2, an overview of basic concepts and related work are depicted. An interworking architecture for NGWN/4G is presented in Section 6.3. The proposed mobility management protocol (eHPIN) is described in Section 6.4. Performance analysis and numerical results are shown in Section 6.5 and 6.6, respectively. Finally, Section 6.7 concludes the paper.

6.2 Background and Related Work

Mobility management enables a system to locate roaming terminals in order to deliver data packets (i.e., *location management*) and to maintain connections with them as they move into a new subnet (i.e., *handoff management*). Handoff management is a major component of mobility management since an MN can trigger several handoffs over a session lifetime as it will be the case in NGWN/4G. It is crucial to provide seamless mobility and service continuity support based on intelligent and efficient techniques. Various schemes have been proposed in the literature and by the IETF for mobility management in IP-based wireless networks.

Mobile IPv6 (MIPv6) (Johnson *et al.* , 2004) was proposed for mobility management at the IP layer and allows MNs to remain reachable in spite of their movements within IP wireless environments. Each MN is always identified by its home address, regardless of its current point of attachment to the network. While away from its home network, an MN is also associated with a care-of address (CoA), which provides information about the MN's current location. After acquisition of CoA, an MN sends a binding update (BU) message to the home agent (HA), informing it of

the new address and also to all active correspondent nodes (CNs) to enable route optimization. However, MIPv6 has some well-known drawbacks such as signaling traffic overhead, high packet loss rate and handoff latency, thereby causing user-perceptible deterioration of real-time traffic (Pérez-Costa *et al.* , 2003; Gwon *et al.* , 2004). Such weaknesses led to the investigation of other solutions designed to enhance MIPv6 and support micro-mobility of MNs.

Two main MIPv6 extensions proposed by the IETF are Hierarchical MIPv6 (HMIPv6) (Soliman *et al.* , 2005) and Fast Handovers for MIPv6 (FMIPv6) (Koodli, 2005). HMIPv6 handles local handoffs through a special node called Mobility Anchor Point (MAP). The MAP, acting as a local HA in the network visited by the MN, limits the amount of MIPv6 signaling outside its domain and reduces delays associated to location update procedure. However, HMIPv6 cannot meet the requirements for delay sensitive traffic, such as voice over IP (VoIP), due to packets loss and handoff latency. FMIPv6 was proposed to reduce handoff latency and to minimize services disruption due to MIPv6 operations during handoffs such as movement detection, binding update and addresses configuration. The link layer information (L2 trigger) is used either to predict or respond rapidly to handoff events.

Although FMIPv6 paves the way on improving MIPv6 performance in terms of handoff latency, it remains hindered by several problems such as QoS support and scalability. In fact, FMIPv6 does not effectively reduce global signaling and packet loss, which cause unacceptable service disruption. In FMIPv6, a new access router (NAR) consumes storage space to buffer the forwarded packets by previous access router (PAR) before delivering packets to the MN. Moreover, these transferred packets lack QoS guarantee before the new QoS path is setup. Combining HMIPv6 and FMIPv6 motivates the design of Fast Handover for HMIPv6 (F-HMIPv6) (Jung *et al.* , 2005a) to allow more network bandwidth usage efficiency.

However, F-HMIPv6 may inherit drawbacks from both FMIPv6 and HMIPv6, such as synchronization issues and signaling traffic overhead that result in combining both schemes (Pérez-Costa *et al.* , 2003; Gwon *et al.* , 2004).

To achieve seamless mobility across various access technologies and networks, an MN needs information about the wireless network to which it could attach. Also, it is necessary to transfer information (context transfer) related to the MN from the current access router to the next one. The Candidate Access Router Discovery (CARD) protocol (Leibsch *et al.* , 2005) and the Context Transfer Protocol (CXTP) (Loughney *et al.* , 2005) have been proposed to enable this procedure. They prevent the use of limited wireless resources, provide fast mobility and secure transfers. Their key objectives consist of reducing latency and packet loss, and avoiding the re-initiation of signaling to and from an MN from the beginning. However, context transfer is not always possible, for example, when an MN moves across different administrative domains. The new network may require the MN to re-authenticate and perform signaling from the beginning rather than to accept the transferred context. Moreover, the entities which exchange context or router identities must authenticate each other. This could become a tedious process in NGWN. All of the aforementioned remarks show that seamless mobility and service continuity are not guaranteed in the current IPv6-based mobility management protocols.

6.3 Interworking Architecture for NGWN

Heterogeneity, in terms of radio access networks in NGWN, requires the integration and interworking of various existing wireless systems. Two majors architectures (*loose* and *tight coupling*) for 3G/WLAN interworking have been proposed by both 3G wireless network initiatives, 3GPP and 3GPP2, for their respective systems (3GPP, 2004; 3GPP2, 2004). However, this integration brings new challenges such

as architecture and protocols design, mobility management, QoS guarantees, interworking and security. All scenarios listed in 3GPP (2004) and 3GPP2 (2006) are not yet fulfilled. Moreover, both interworking models have as well as pros and cons.

An interworking architecture, called *Integrated InterSystem Architecture* (IISA) based on 3GPP/3GPP2-WLAN interworking models, was proposed in Makaya & Pierre (2007b) and shown in Fig. 6.1. For the sake of simplicity, only UMTS, CDMA2000 and WLAN networks are illustrated. However, IISA may integrate any number of radio access technologies (RATs) and mobile devices may be equipped with any number of interfaces. Instead of developing new infrastructures, IISA extends existing infrastructures to tackle integration and interworking issues and provide mobile users with ubiquity or *always best connected.* The serving GPRS (general packet radio service) support node (SGSN) and packet control function (PCF) are enhanced with the AR functionalities and called *Access Edge Node* (AEN). Similarly, the gateway GPRS support node (GGSN) and packet data serving node (PDSN) are extended with the MAP or HA functionalities (to enable message format conversion, QoS requirements mapping, etc.) and called *Border Edge Node* (BEN). The WLAN Interworking Gateway (WIG) acts as a route policy element and ensures message format conversion.

A novel entity, *Interworking Decision Engine* (IDE), is introduced to enable the interworking and handoffs between various networks by reducing signaling traffic, services disruption during handoff and handles authentication, authorization and accounting (AAA) and mobility management. The usage of the IDE could be considered as a value-added service that network operators offer to their subscribers to allow roaming in other networks. To avoid additional signaling overhead due to the execution of the AAA procedure every time an MN performs handoff and requests registration, we propose a token-based approach. The token includes security association parameters to setup secure tunnel between an MN and AR/AENs.

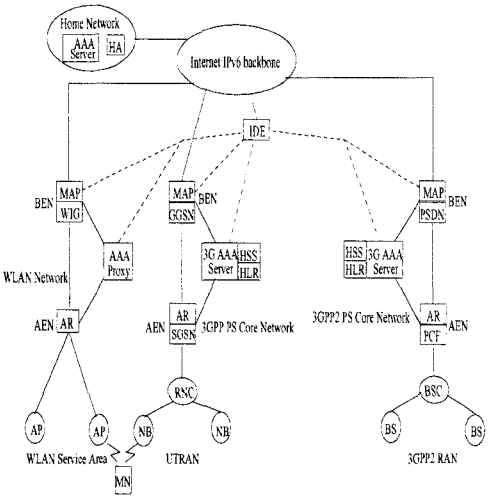

Figure 6.1 Integrated InterSystem Architecture (IISA).

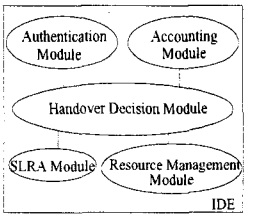

Figure 6.2 Interworking Decision Engine (IDE).

The logical components of the IDE are illustrated in Fig. 6.2. The *Authentication Module* (AuM) is used to authenticate users moving across different wireless networks and avoids the required direct security agreements or association between foreign networks and home network. The AuM stores information such as the subscribers' identities, users' preferences/profiles and terminal mobility patterns. The *Accounting Module* (AcM) enables billing between different wireless networks and stores billing information associated with the resource usage. It acts as common billing/charging system between various network operators.

Usually, different administrative domains have different QoS policies for resources allocation. Thus, when an MN moves from one administrative domain to another, QoS re-negotiation may be required. Such re-negotiation will be based on service level agreements (SLAs) between both domains. The *Resource Management Module* (RmM) enables QoS mapping and re-negotiation. Furthermore, the RmM allows fast transfer of user profiles and QoS parameters between two administrative domains during handoff. The *SLRA Module* stores information about service pro-

viders or network operators which have SLAs and roaming agreements (RAs) with the IDE manager. The *Handover Decision Module* (HdM) is used when intersystem or inter-domain handoff should be granted or not. In other words, it provides support for roaming and handoffs. For further details about the IISA architecture, please refer to Makaya & Pierre (2007b).

6.4 Proposed eHPIN Protocol

Assuming that mobile devices are becoming increasingly powerful, intelligent and sensitive to link layer changes, a network-assisted and mobile-controlled handoff strategy is adopted. eHPIN combines both mobile-monitored and network-probed information to provide reliable handoff control. Prior to handoff, an MN can obtain information regarding candidate wireless networks to which it is likely to handoff and uses such information to optimize handoff performance. On the other hand, if mobile device capabilities are limited, handoff decision is taken by mobility agents in the network side.

L2 trigger generation may be imprecise because it is a link layer event and depends on L2 technology and channel conditions. Thus, two modes (predictive and reactive) have been proposed for FMIPv6. When coupled with CARD protocol, FMIPv6 can be inefficient. Hence, it is necessary to find an effective way to perform access router discovery procedure and handoff anticipation in a one-suite protocol. eHPIN is proposed to reach this goal. In eHPIN, the handshake procedure for access router discovery and fast handoff as well as all time consuming operations such as bi-directional tunnels setup between the MAP/BEN and candidate access routers (CARs) or AENs, duplicate address detection (DAD) procedure and CAR/AEN pre-selection are performed before the L2 trigger generation.

6.4.1 Handoff Initiation with eHPIN

With the information exchanged between the MAP/BEN and AR/AENs by using *Router Information eXchange* (RIX Request/Reply) messages, the BEN maintains a global view (i.e., load status of AENs, connection state of any MN in its domain as well as movement patterns) of its domain and can learn both L2 and L3 information of an access network. L2 information may include the specific link layer wireless access technology, system parameters (e.g., channel frequency and number) and QoS status such as bandwidth availability and signal strength. L3 information can include the global address of AR/AEN, the address of the prefix advertised in the wireless network, the current QoS status and parameters. The QoS parameters may include information such as the supported data rate, the video coding rate, and maximal delay. L2 and L3 information are then forwarded to the IDE and allows it to maintain a global view of all MAP/BEN domains having SLAs with the IDE manager. The exchange of RIX messages is quite similar as that of the routing information protocol (RIP) (Hedrick, 1988) works to allow neighboring routers to exchange their routing table with one another. The update interval time (e.g., 30 seconds) for each information depends on its property : static or dynamic. Thus, the backbone signaling increment does not require high additional costs for system deployment.

The MN decides whether to send the CARD Request message to MAP/BEN according to the generation of the anticipated triggers (AT). For example, high bit error rate, link going down and weak signal strength, security risks, monetary cost and geographical location can be used as anticipated triggers. The CARD Request message contains user preferences as well as information regarding the applications required QoS capabilities. To allow seamless service continuity, the requirements specified in the CARD Request message need to be set consistently with

the QoS negotiated in the previous subnet. Crucial for real-time applications, QoS consistency is handled by the IDE, which allows QoS mapping between different networks. Upon receipt of the CARD Request message, the MAP/BEN checks its local CAR/AEN table to retrieve information about CAR/AENs' capabilities. Moreover, the MAP/BEN performs pre-filtering of available AR/AENs in order to have potential CAR/AENs, address auto-configuration (AA) process on the behalf of an MN in order to form one or more new on-link CoAs (NLCoAs). For address auto-configuration, we assume that the new CoAs pool is located at the MAP/BEN and which is updated by an out-of-band signaling based on RIX message exchanged between the MAP/BEN and AR/AENs. The MAP/BEN relieves the MN of the burden of LCoAs and RCoAs computation.

Note that if the MAP/BEN lacks information regarding this user profile, it requests such data to the IDE rather than to the MN's HA, which is likely to be far away from the current location. After receiving the CARD Request message, the MAP/BEN sends a handoff initiate (HI) message containing the corresponding NLCoAs to the potential CAR/AENs. When all potential CAR/AEN receive the HI message containing NLCoA, they perform a duplicate address detection (DAD) procedure and acts as a proxy for the MN to defend this temporary address in its network. The HI is also used to trigger the request of context transfer. In other words, the MAP/BEN transmits a Context Transfer Data (CTD) message, piggybacked in HI, to CAR/AENs. Example of features contained in CTD message are QoS context information, header compression, security details, authentication, authorization and accounting (AAA) information. This paper focuses mainly on QoS context information. Performing a DAD procedure for all possible NLCoAs with eHPIN requires some extra overhead compared to basic F-HMIPv6 and FMIPv6. However, the DAD procedure is performed prior to the L2 trigger generation, then it reduces L3 handoff latency and the impacts of imprecise L2 trigger timing.

When the CAR/AEN receives a CTD message, it may generate a CTD Reply (CTDR) message optionally to report the status of processing the received contexts and this message is piggybacked in the handoff acknowledgment (HAck) message. The CAR/AEN installs the contexts once it is received from the MAP/BEN. This context will be activate upon receiving a fast binding update acknowledgment (FBAck) message. The CAR/AEN will send a HAck message to the MAP/BEN only after relocation of traffic bearers and resources are reserved for the new path in order to indicate that handoff may be done and packets forwarding may be initiated. The MAP/BEN binds previous on-link CoAs (PLCoAs) and the NLCoA, but marks its state idle and sends a CARD Reply message to the MN which contains the NLCoAs set, CAR/AENs list and capabilities. The idle state means that, the MAP/BEN does not start buffering and forwarding packets at this stage, nor does it uses reserved resources for this handoff preparation request. Contrary to FMIPv6 and F-HMIPv6, where forwarded packets lack QoS guarantees before the new QoS path is set up, eHPIN solves this issue.

With the CARD Request/Reply messages exchange, an MN knows the candidate AENs to which it is likely to handoff. Then, the MN will activate only the interface associated to the CAR/AEN list, rather than setting all air-interfaces always on as is the case with traditional IPv6-based mobility management schemes. This selective interface activation enables better trade-off between system discovery time and power consumption efficiency. After receiving the CARD Reply, an MN can start a handoff any time. The CARD Request/Reply messages exchange no longer delays the handoff procedure, as it is carried out while the MN uses the previous on-link CoA (PLCoA). Whenever an MN receives the L2 trigger, it initiates a target AR/AEN selection among the CAR/AENs set. This selection is based on the handoff decision function proposed in Makaya & Pierre (2006).

6.4.2 Handoff Execution with eHPIN

Once the handoff decision step is completed the MN sends a fast binding update (FBU) message containing the selected target AR/AEN information to the MAP/BEN. Unlike the basic fast handoff schemes, the FBU message is not used to trigger bi-directional tunnel establishment or handoff initiate/acknowledgment (HI/HAck) messages exchanges, but rather triggers the packet forwarding procedure. Upon receipt of the FBU message, the MAP/BEN activates the idle binding, sends the fast binding update acknowledgment (FBAck) message to the MN on both links (previous and new) and establishes a binding between PLCoA and NLCoA. The MAP/BEN can start packets forwarding to the target NAR/AEN.

6.4.2.1 Intra-BEN Roaming Scenario

When the selected NAR/AEN among CAR/AENs receives the FBAck message, it activates the transferred context. The MN performs a L2 handoff and sends fast neighbor advertisement (FNA) message to announce its presence on the new link. Upon receiving the FNA, the NAR/AEN starts to deliver the buffered packets, if any, to the MN. The disordering packets problem can be reduced significantly with buffering at the NAR/AEN or MN. In fact, a routing header extension is added to the forwarding packets before they are forwarded to NAR/AEN. The routing header extension contains the previous LCoA while the source address in the IPv6 header has the HA/CN address. Hence, the MN can differentiate forwarded packets from regular packets from HA/CNs. The MN does not deliver packets from HA/CN to its upper layer before all forwarded packets are delivered.

The binding update (BU) procedure is performed by NAR/AEN on the behalf of the MN. In fact, the AR/AEN acts as a proxy and copies a BU list of the MN in

its cache and manages this list (e.g., lifetime entries) in the same way as the original is managed by the MN. The AR/AEN cache copy must be periodically updated in accordance with the original BU list of the MN. As soon as an MN becomes attached to NAR/AEN, the copy of the BU list is used by the NAR/AEN to inform the MAP/BEN about the NLCoA. When the lifetime of the BU list cached in the AR/AEN is about to expire, the AR/AEN can send a BU list renewal request to the MN. The BU list renewal is conducted in the same way as a basic BU refresh Johnson *et al.* (2004). The MN sends the BU list to NAR/AEN and simultaneously it attaches to the NAR/AEN. Piggybacking the BU list in a signaling message prevents separate out-of-band messages from MN to NAR/AEN, thus, reducing the signaling traffic overhead.

Note that, if the FBU message was not sent before the L2 handoff, then an MN sends it piggybacked in a FNA message (FNA[FBU]) over the new link. When the NAR/AEN receives a FNA[FBU] message, it processes the FNA message part, extracts the FBU message and forwards it to the serving MAP/BEN. When the MAP/BEN receives the FBU message, it responds by sending FBAck message to NAR/AEN. At this time, the MAP/BEN can start tunneling incoming and in-flight packets addressed to PLCoA towards NLCoA. This procedure refers to the reactive mode of eHPIN while the predictive mode is explained above (i.e., the MN sends FBU message through PAR/AEN's link and FBAck is received before the L2 handoff). The reactive mode is carried out either intentionally or serve as a fall-back mechanism when the predictive mode cannot be completed successfully, for example, if the L2 handoff is completed before the FBAck message is received at the MN. Signaling messages exchange of eHPIN is shown in Fig. 6.3 for intrasystem or intersystem handoff for intra-MAP/BEN roaming. Contrary to basic fast handoff schemes (i.e., FMIPv6, F-HMIPv6), only one round trip message exchange for FBU/FBAck and FNA are required for handoff after L2 trigger with eHPIN.

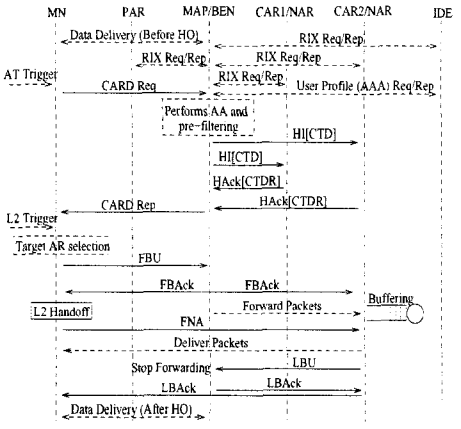

Figure 6.3 Signaling messages with eH-PIN for intra-BEN roaming.

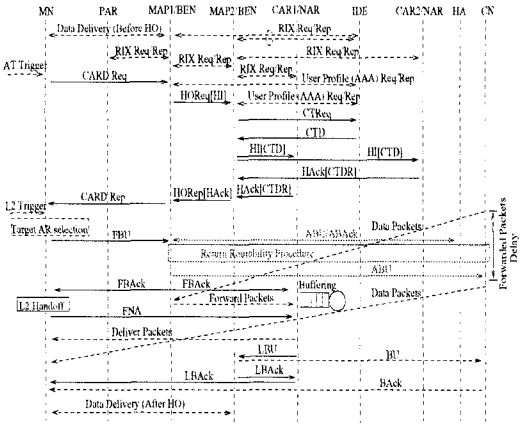

Figure 6.4 Signaling messages with eH-PIN for inter-BEN roaming.

6.4.2.2 Inter-BEN Roaming Scenario

If the CAR/AENs are located within another MAP/BEN domain, the serving MAP/BEN (MAP1/BEN in Fig. 6.4) sends a handoff request (HOReq) message to the candidate MAP/BEN (MAP2/BEN in Fig. 6.4) and encapsulates a HI message within HOReq. The candidate MAP/BEN forwards HI message virtually in parallel to all CAR/AENs belonging to its domain by including the CTD message. Note that, if the context information of this MN are not available at the candidate MAP/BEN, the latter sends Context Transfer Request (CTReq) message to the IDE in order to obtain the session management parameters of the MN for establishment of traffic bearers on the new path. In response to a CTReq message, the IDE transmits a CTD message that includes the MN's feature contexts. When the new MAP/BEN receives a CTD message from the IDE, it installs the contexts.

Once the application requirements are validated, the CAR/AENs send a HAck message to the candidate MAP/BEN, which then encapsulates a HAck message with the handoff reply (HORep) message and sends it to the current MAP/BEN. HORep[HAck] contains NLCoAs, CAR/AENs capabilities and other adequate in-

formation. Upon receipt of HORep[HAck] message, the serving MAP/BEN sends a CARD Reply message including the CAR/AENs list and capabilities, associated NLCoAs and other information. Similar operations as those in the case of intra-MAP/BEN roaming ensue when a L2 trigger is generated. Fig. 6.4 shows the message sequence exchange for eHPIN for intrasystem and/or intersystem handoff between two MAP/BEN domains (inter-MAP/BEN handoff).

In fast handoff schemes, forwarded packets experience additional delays due to the buffering at the NAR/AEN and the sub-optimal route they took. To reduce this forwarding delay, eHPIN allows an MN to inform active CNs about its new RCoA immediately after its validation. In fact, during inter-MAP/BEN roaming, when the previous MAP/BEN receives a FBU message, it can immediately send an anticipated BU (ABU) message on the behalf of an MN to the HA in order to notify it to perform anticipated binding update. Moreover, the MAP/BEN performs the return routability procedure (RR) (Johnson *et al.* , 2004) and anticipated BU to all active CNs with the help of the cached BU list on the behalf of the MN. Hence, the data routing will be conducted early through optimal path between CNs and MN. Thus, the number of packets forwarded (between the previous MAP/BEN and NAR/AEN) and their delay is minimized. This alleviates one of the major drawbacks of basic fast handoff schemes, i.e., the packet delay introduced by the tunneling procedure.

6.5 Performance Evaluation

In IP-based wireless networks, QoS may be defined in terms of packets loss, handoff latency and signaling traffic overhead cost. Analysis of these metrics are very useful to evaluate the performance of mobility management protocols. The notation used in this paper is outlined in Table 6.1. User mobility and traffic models

Tableau 6.1 Notation.

C^g	global binding update cost to HA/CNs
C^l	local binding update cost to MAP/BEN
M	number of subnets in the MAP/BEN domain
N_{CN}	number of CNs with a binding cache entry for an MN
$d_{X,Y}$	average number of hops between nodes X and Y
$C_{X,Y}$	transmission cost of control packets between nodes X and Y
PC_X	processing cost of control packet at node X
t_T	time period between the L2 trigger and the start of the L2 switch
t_F	time period between the transmission of a FBU and the start of the L2 switch
P_s	success probability of the anticipated handoff

are crucial for efficient system design and performance evaluation. Usually, MN mobility is modeled by the cell residence time and numerous random variable types are used for this purpose Fang (2003). We consider a traffic model composed of two levels, session and packets. The session duration follows exponential distribution with the inter-session rate λ_s while the packet generation and arrival rate follow a Poisson process. The evaluation of time that an MN stays within the subnet is usually based on two distributions : Gamma and Exponential.

The Gamma distribution is very realistic for mobility model by considering changes in the speed and direction of the MN while the Exponential distribution is a particular case of Gamma distribution. We consider that subnet and MAP/BEN domain residence time follow Gamma distribution with a mean of $1/\mu_c$ and $1/\mu_d$, respectively. Note that, μ_c is the border crossing rate for an MN moving out of an AR/AEN coverage area and μ_d for an MN moving out of the MAP/BEN domain. When an MN crosses the MAP/BEN domain border, it also crosses an AR/AEN border. Hence, the rate for AR/AEN crossing for which the MN remains in the MAP/BEN domain is $\mu_l = \mu_c - \mu_d$. If we assume that all subnets are made up of circular shapes forming together a contiguous area and that each MAP/BEN

domain is composed of M equally subnets, we obtain $\mu_d = \frac{\mu_c}{\sqrt{M}}$.

6.5.1 Total Signaling Cost

The performance analysis of wireless networks must consider the total signaling cost induced by mobility management schemes. As for wireless cellular networks, signaling traffic overhead cost must be performed for NGWN or IP-based mobile environments. In NGWN, there are two kinds of location or binding update signaling. One takes place from an MN's subnet crossing and another occurs when the binding is about to expire. Moreover, delivery of data packets induces usage of network resources, thus generating additional costs. Hence, the total signaling cost, C_T, could be divided into the binding update signaling cost, C_{BU}, and the packet delivery cost, C_{PD} : $C_T = C_{BU} + C_{PD}$.

6.5.1.1 Binding Update Signaling Cost

The binding update cost heavily depends on the average number of location updates during the inter-session arrival time. Depending on the type of movement, two kinds of location or binding updates could be performed : local and global binding update. The global binding update procedure refers to the registration of RCoA to HA/CNs. On the other hand, if an MN changes its current address (LCoA) within a MAP/BEN domain, it only needs to register this new LCoA to the MAP/BEN. Hence, the average location binding update cost for IPv6-based mobility management schemes during inter-session time can be expressed by :

$$C_{BU} = E(N_l)C^l + E(N_d)C^g \tag{6.1}$$

where $E(N_l)$ is the average number of subnets (AR/AENs) that an MN crosses while remaining within a given MAP/BEN domain during an ongoing session and $E(N_d)$ denotes the average number of MAP/BEN domains crossing.

To perform a signaling overhead analysis, a performance factor called session-to-mobility ratio (SMR) is introduced. It is similar to the call-to-mobility ratio (CMR) defined in cellular networks (Xie & Akyildiz, 2002). The SMR represents the relative ratio of session arrival rate over the user mobility rate : $SMR = \lambda_s/\mu_c$. The binding update signaling cost, C_{BU}, is then given by :

$$C_{BU} = \frac{1}{\lambda_s}\left(\mu_d C^g + \mu_l C^l\right) = \frac{1}{SMR\sqrt{M}}\left[C^g + (\sqrt{M}-1)C^l\right]. \tag{6.2}$$

In IP-based mobile environments, not all L2 handoffs result in L3 handoffs. Hence, handoff procedure anticipated by using L2 trigger may lead to unnecessary signaling traffic. The critical phase of the fast handoff approach starts when a L2 trigger is generated to indicate the impending handoff. We assume that if an MN receives a FBAck message from the MAP/BEN, that it will inevitably start L3 handoff to the NAR/AEN without exceptions. Hence, if there is no real handoff after a L2 trigger generation, all messages exchanged from FBU to FBAck may be unnecessary. The global and local binding update signaling cost for eHPIN are expressed as follows :

$$\begin{aligned} C^g &= P_s S_s^g + (1-P_s)(S_f^g + S_r^g) + C_{ru} \\ C^l &= P_s S_s^l + (1-P_s)(S_f^l + S_r^l) + C_{mu} \end{aligned} \tag{6.3}$$

where C_{ru} represents the binding update cost at the HA/CNs, C_{mu} the binding update cost at the MAP/BEN, S_s^g (resp. S_s^l) the global (resp. local) signaling cost for a successfully anticipated handoff, S_f^g (resp. S_f^l) the global (resp. local) signaling cost for control messages if no real L3 handoff occurs and S_r^g (resp. S_r^l) the global

(resp. local) signaling cost for the reactive mode. Table 6.2 shows their expressions.

Tableau 6.2 Expression of partial signaling costs.

C_{mu} =	$2C_{AEN,BEN} + PC_{BEN}$
C_{ru} =	$2(C_{BEN,HA} + N_{CN}C_{BEN,CN}) + PC_{HA} + N_{CN}PC_{CN}$
S_f^l =	$C_{MN,BEN} + C_{BEN,AEN} + PC_{BEN}$
S_s^l =	$2C_{MN,BEN} + C_{BEN,AEN} + C_{MN,AEN} + PC_{BEN} + PC_{AEN}$
S_r^l =	$C_{MN,AEN} + 2C_{AEN,BEN} + PC_{BEN} + 2PC_{AEN}$
S_f^g =	$C_{MN,pBEN} + (C_{pBEN,nBEN} + C_{nBEN,AEN}) + PC_{BEN}$
S_r^g =	$C_{MN,AEN} + 2(C_{AEN,nBEN} + C_{nBEN,pBEN}) + 2PC_{BEN} + PC_{AEN}$
S_s^g =	$2C_{MN,pBEN} + (C_{pBEN,nBEN} + C_{nBEN,AEN}) + C_{MN,AEN} + 2PC_{BEN} + PC_{AEN}$

The packet transmission cost in IP-based networks is proportional to the distance in hops between the source and the destination nodes. Furthermore, the transmission cost in a wireless link is generally superior than that in a wired link (Xie & Akyildiz, 2002). Thus, the transmission cost of a control packet between nodes X and Y belonging to the wired part of a network can be expressed as $C_{X,Y} = \tau d_{X,Y}$ while $C_{MN,AEN} = \tau\kappa$, where τ is the unit transmission cost over wired link and κ the weighting factor for the wireless link.

6.5.1.2 Packet Delivery Cost

Similarly to Koodli & Perkins (2001), we divide handoff latency into three components : link switching or L2 handoff latency (t_{L2}), IP connectivity latency (t_{IP}) due to movement detection and address configuration and location update latency (t_U). The IP connectivity latency reflects how quickly an MN can send IP packets after the L2 handoff, while the location update latency represents the delay requi-

red for forwarding IP packets to MN's new IP address. On the other hand, the time period between the starting point of L2 handoff and the moment an MN receives IP packets for the first time through new link refers to packet reception latency (t_P) or data latency. Moreover, the following delay components are introduced : movement detection delay (t_{MD}), address configuration and DAD procedure delay (t_{AC}), binding update latency (t_{BU}) and delay from completion of binding update and reception of the first packet by an MN through the new IP address (t_{NR}).

The timing diagram of eHPIN for intra-MAP/BEN roaming is illustrated in Fig. 6.5. When two endpoints have an ongoing session, a packet delivery cost incurs.

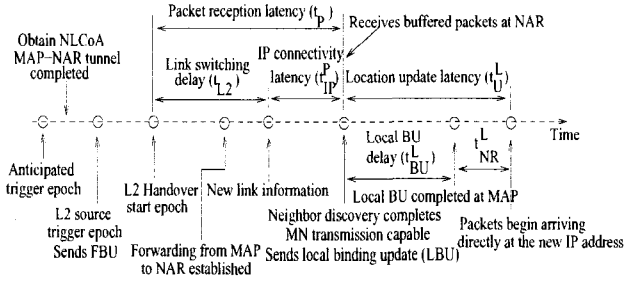

Figure 6.5 Handoff delay timeline of eHPIN for intra-BEN roaming.

The packet delivery cost consists of the packet transmission cost and the packet processing cost. By using the handoff timing diagram illustrated in Fig. 6.5, the packet delivery cost could be defined as the linear combination of the packet tunneling/forwarding cost (C_{tun}) and the packet loss cost (C_{loss}). Let α and β be the weighting factors (where $\alpha + \beta = 1$), which emphasize the tunneling and dropping effects. The packet delivery cost, C_{PD}, is computed as follows :

$$C_{PD} = \alpha C_{tun} + \beta C_{loss}. \tag{6.4}$$

Let s_c and s_d be the average size of control and data packets, respectively and $\eta = s_d/s_c$. The cost of transferring data packets is η greater than the cost of

transmitting control packets. Let λ_p be the packet arrival rate in unit of packet per time. The packet loss in fast handoff schemes may be due either to L2 handoff or in case of wrong spatial prediction of NAR/AEN. The packet loss due to L2 handoff delay is inevitable without efficient buffering mechanisms (Koodli & Perkins, 2001). Since a bi-directional tunnel is established before L2 trigger, there is no packet loss cost for the predictive mode of eHPIN (i.e., $C_{loss}^{p,l} = 0$). Moreover, as the packets forwarding process is not supported in the reactive mode, packet tunneling cost equal zero ($C_{tun}^{r,l} = 0$). Due to wrong spatial prediction of NAR/AEN or if FBAck message was not received through the previous link, the packets forwarded by the MAP/BEN to an erroneously predicted NAR/AEN can be lost. Packets forwarding to the wrong NAR/AEN stops when the FBU message sent through the NAR/AEN's link is received at the MAP/BEN. In this case, the reactive mode of eHPIN is used.

The packet tunneling cost for predictive mode ($C_{tun}^{p,l}$) and the packet loss cost ($C_{loss}^{r,l}$) of eHPIN are expressed as follows :

$$C_{tun}^{p,l} = \lambda_p C_{cm}^{s,l}(t_{L2} + t_{IP}^{P} + t_{U}^{L}) \quad \text{and} \quad C_{loss}^{r,l} = \lambda_p C_{cm}^{r,l}(t_{L2} + t_{IP}^{R} + t_{U}^{L}) \tag{6.5}$$

where $t_U^L = t_{BU}^L + t_{NR}^L$ is the location update latency for intra-MAP/BEN movement, t_{IP}^R is the IP connectivity latency of reactive mode, t_{IP}^P is the IP connectivity latency excluding IP addresses configuration, DAD procedure and movement detection. In fact, these operations are performed in anticipation prior an MN leaves the PAR/AEN's link. The cost of transferring data packets from an active CN to MN through to the MAP/BEN is $C_{cm}^{s,l} = \eta(C_{CN,BEN} + C_{BEN,AEN} + C_{AEN,MN})$ and $C_{cm}^{r,l} = \eta(C_{CN,BEN} + C_{BEN,AEN} + C_{AEN,MN})$. The average packet delivery cost of eHPIN is given by :

$$C_{PD}^{a,l} = P_s C_{PD}^{p,l} + (1 - P_s) C_{PD}^{r,l} \tag{6.6}$$

where $C_{PD}^{p,l}$ and $C_{PD}^{r,l}$ indicate packet delivery costs for the predictive and reactive modes of eHPIN, respectively, and are computed using (6.4).

On the other hand, for inter-MAP/BEN roaming case with eHPIN, the timing diagram is illustrated in Fig. 6.6, where t_{HA} is the delay to perform anticipated BU or to register a new RCoA to the HA, t_{RR} is the delay for the return routability procedure and t_{CN} represents the delay for performing anticipated BU or registering a new RCoA to all active CNs. The packet loss cost (C_{loss}) and the packet tunneling

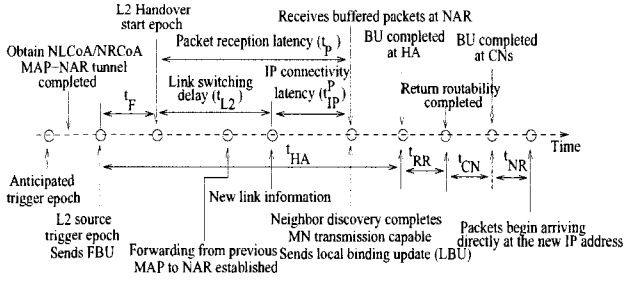

Figure 6.6 Handoff delay timeline of eHPIN for inter-BEN roaming.

cost (C_{tun}) are expressed as follows :

$$\begin{aligned} C_{loss}^{r,g} &= \lambda_p C_{cm}^{r,g}(t_{L2} + t_{IP}^{RG} + t_U) \\ C_{tun}^{p,g} &= \lambda_p C_{cm}^{s,g}[\max(t_{L2} + t_{IP}^{P}, t_{HA} - t_F) + t_{RR} + t_{CN} + t_{NR}] \end{aligned} \tag{6.7}$$

where $t_U = t_{BU} + t_{RR} + t_{NR}$, t_{IP}^{RG} is the IP connectivity latency of reactive mode for inter-MAP/BEN roaming, the cost of transferring data packets from an active CN to MN by transiting to the previous and the new MAP/BEN is $C_{cm}^{s,g} = \eta(C_{CN,pBEN} + C_{pBEN,nBEN} + C_{nBEN,AEN} + C_{AEN,MN})$ and $C_{cm}^{r,g} = \eta(C_{CN,pBEN} + C_{pBEN,PAR} + C_{AEN,MN})$. The average packet delivery cost for eHPIN associated to inter-MAP/BEN roaming is computed similarly as in (6.6) and by using (6.4). eHPIN eliminates all sources of packet loss except for the unavoidable loss due to the link layer switching handoff. However, with efficient buffering mechanism at AR/AENs packet loss during L2 handoff may be avoided.

6.5.2 Handoff Latency and Packet Loss

Handoff latency and packet loss are computed according to the following parameters : t_{L2} represents the L2 handoff latency and $t_{X,Y}$ one-way transmission delay of a message of size s between nodes X and Y. If one of the endpoints is an MN, $t_{X,Y}$ is computed as follows :

$$t_{X,Y}(s) = \frac{1-q}{1+q}\left(\frac{s}{B_{wl}} + L_{wl}\right) + (d_{X,Y} - 1)\left(\frac{s}{B_w} + L_w + \varpi_q\right) \tag{6.8}$$

where q is the probability of wireless link failure, ϖ_q the average queueing delay at each router on the Internet (McNair *et al.* , 2001), B_{wl} (resp. B_w) the bandwidth of the wireless (resp. wired) link and L_{wl} (resp. L_w) wireless (resp. wired) link delay.

Let Δ_{ns} be the time elapsed between the reception of FBAck on the previous link and the beginning of the L2 handoff when there is no good synchronization between L2 and L3 handoff operations. Moreover, let Δ_{lr} be the time between the last packet received through the previous link and the L2 handoff beginning when the FBAck arrives on the new link. Note that, Δ_{lr} and Δ_{ns} can equal zero. For eHPIN, handoff latency depends on the information available, and on which link fast handoff messages are exchanged. If information regarding the NAR/AEN and impending handoff are available and, if the FBAck message is received through the previous link, the handoff latency is expressed as follows :

$$O^l_{eHPIN} = \Delta_{ns} + t_{L2} + 2t_{MN,AEN}. \tag{6.9}$$

However, if a FBAck message is not received on the previous link, it will be received through the new link. Hence, in this case, the handoff latency for eHPIN is expressed as follows :

$$N^l_{eHPIN} = \Delta_{lr} + t_{L2} + 2t_{MN,AEN} + 3t_{AEN,BEN}. \tag{6.10}$$

The average handoff latency of eHPIN for intra-MAP/BEN roaming is given by :

$$D^{l}_{eHPIN} = P_s O^{l}_{eHPIN} + (1 - P_s) N^{l}_{eHPIN}. \tag{6.11}$$

For inter-MAP/BEN movement, when FBAck message is received through the previous link, the handoff latency of eHPIN is identical as for intra-MAP/BEN roaming : $O^{g}_{eHPIN} = O^{l}_{eHPIN}$. In fact, the handoff procedure depends only on intra-MAP/BEN communication delay, since the inter-MAP/BEN signaling is completed before the L2 handoff. On the other hand, when a FBAck message is received through the new link for inter-MAP/BEN movement, it is assumed that appropriate information about NAR/AEN is already available and NLCoA is already configured. Hence, the handoff latency of eHPIN for inter-MAP/BEN roaming is given by :

$$N^{g}_{eHPIN} = \Delta_{lr} + t_{L2} + 2t_{MN,AEN} + 3[t_{AEN,nBEN} + t_{nBEN,pBEN}]. \tag{6.12}$$

The average handoff latency of eHPIN for inter-MAP/BEN roaming is computed similarly as in (6.11).

In theory with eHPIN, there are no packets loss, unless buffers overflow at NAR/AEN or MAP/BEN. However, without efficient buffer management, packets forwarded can be lost during handoff latency. In fact, the number of packets lost is proportional to handoff latency :

$$P^{eHPIN,l}_{loss} = \begin{cases} \max(BS^{l}_{eHPIN} - B, 0) & \text{for efficient buffer management} \\ \lambda_p D^{l}_{eHPIN} & \text{otherwise} \end{cases} \tag{6.13}$$

where B is the buffer size of an AR/AEN and BS^{l}_{eHPIN} is the required buffer space at NAR/AEN for intra-MAP/BEN roaming with eHPIN during packets forwarding

Tableau 6.3 Performance analysis parameters.

Parameters	Symbols	Values
L2 handoff time	t_{L2}	50 ms
Time period between L2 trigger and L2 handoff	t_T	10 ms
Prediction probability	P_s	0.98
Wireless link failure probability	q	0.50
Wired link bandwidth	B_w	100 Mbps
Wireless link bandwidth	B_{wl}	11 Mbps
Wired link delay	L_w	2 ms
Wireless link delay	L_{wl}	10 ms
Number of ARs by domain	M	2
Control packet size	s_c	96 bytes
Data packet size	s_d	200 bytes
Packet arrival rate	λ_p	10 packets/s

and is computed as follows :

$$BS^{l}_{eHPIN} = \lambda_p[P_s(t_{L2} + t^{P}_{IP} + t^{L}_{U}) + (1 - P_s)t^{L}_{NR}]. \tag{6.14}$$

Similarly we can compute the number of packets lost ($P^{eHPIN,g}_{loss}$) and the required buffer space at NAR/AEN (BS^{g}_{eHPIN}) for inter-MAP/BEN roaming.

6.6 Numerical Results

The parameter and default values used in performance evaluation are given in Table 6.3, except when the wireless link delay (L_{wl}), packet arrival rate (λ_p) and prediction probability (P_s) are considered variable parameters. An analytical framework for the performance evaluation of IPv6-based handoff schemes proposed by the IETF, i.e., MIPv6, HMIPv6, FMIPv6 and F-HMIPv6 is presented in Makaya & Pierre (2007a). These evaluation methods are used to compare the performance of the IETF's protocols and that of the eHPIN. The network topology considered

for this analysis is illustrated in Fig. 6.7. We assume that distance between different domains are equal, i.e., $c = d = e = f = 10$ and $a = 1$, $b = 2$, and $g = 4$. All links are considered to be full-duplex in terms of capacity and delay. Other parameters used to compute signaling costs are defined as follows : $\tau = 1$, $\kappa = 10$, $\alpha = 0.2$, $\beta = 0.8$, $PC_{AEN} = 8$, $PC_{HA} = 24$, $PC_{CN} = 4$, $PC_{IDE} = 15$ and $PC_{BEN} = 12$. Performance analysis is conducted using MATLAB and OPNET softwares.

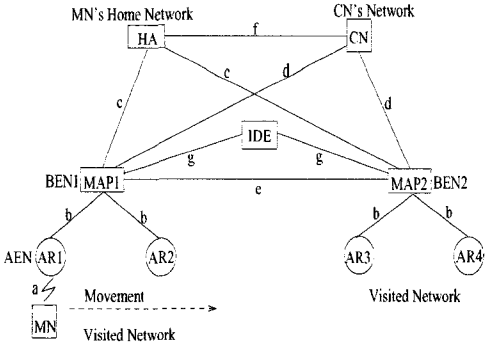

Figure 6.7 Network topology used for analysis.

Fig. 6.8 illustrates the binding update signaling cost during handoff as a function of the SMR. When the SMR is small, the mobility rate is larger than the session arrival rate, the MN changes frequently its point of attachment resulting in several handoffs. These handoffs will cause the exchange of several messages between different entities and will increase signaling overhead. However, when the session arrival rate is larger than the mobility rate (i.e., SMR is larger than 1), the binding update is less often performed. In other words, the signaling overhead decreases as the frequency of the subnet change decreases. eHPIN allows significant signaling overhead cost saving compared to other protocols. The RIX messages exchange introduces additional signaling similarly as with routing information protocol (RIP). However, this signaling increment only occurs in the wired part of network. Compared to the wireless part, the wired one has much superior bandwidth and resources.

Fig. 6.9 illustrates the binding update signaling cost during handoff as a function of the prediction probability (P_s) when $SMR = 0.1$. HMIPv6 and MIPv6 are not affected by the prediction probability contrary to fast handoff-based schemes since they do not use L2 trigger to anticipate the handoff. The signaling overhead decreases when the prediction probability accuracy increases for fast handoff-based schemes (i.e., FMIPv6, F-HMIPv6 and eHPIN). For small values of P_s, HMIPv6 performs better than eHPIN. However, when P_s increases, eHPIN outperforms all other schemes.

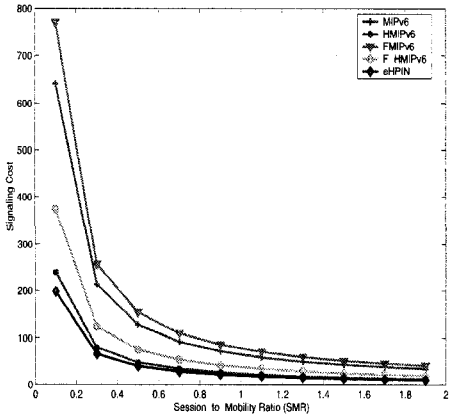

Figure 6.8 Impact of SMR on binding update signaling cost.

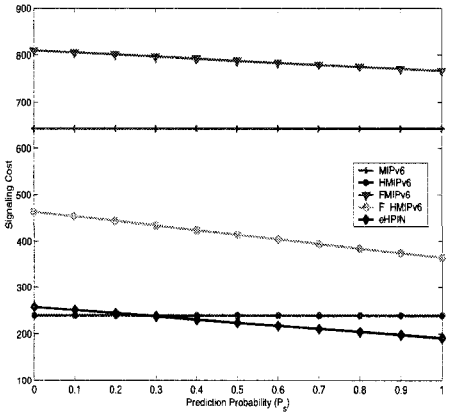

Figure 6.9 Impact of probability P_s on binding update signaling.

The packet delivery cost is shown in Fig. 6.10 as a function of the packet arrival rate (λ_s). Combined hierarchical and fast handoff based schemes (i.e., F-HMIPv6 and eHPIN) perform better than FMIPv6, MIPv6 and HMIPv6. Moreover, they are more efficient when λ_s increases. This means that eHPIN and F-HMIPv6 are more adequate for real-time applications where periodic packets are sent at high rates. We observe that eHPIN enables lower packet delivery cost compared to F-HMIPv6. For varying prediction probability (P_s), Fig. 6.11 shows the packet delivery cost which decreases when the accuracy of P_s increases for fast handoff schemes. The high value of P_s means that the FBAck message is received through the previous link

(i.e., via PAR/AEN). Then, buffered packets are delivered to an MN just after it attaches to the new link. Hence, service disruption delay is reduced. We observe that, regardless of the prediction probability value, eHPIN outperforms all other schemes by providing a lower packet delivery cost. The prediction probability has a greater effect on F-HMIPv6. In fact, when $P_s = 0$, F-HMIPv6 turns to HMIPv6, which is its reactive mode.

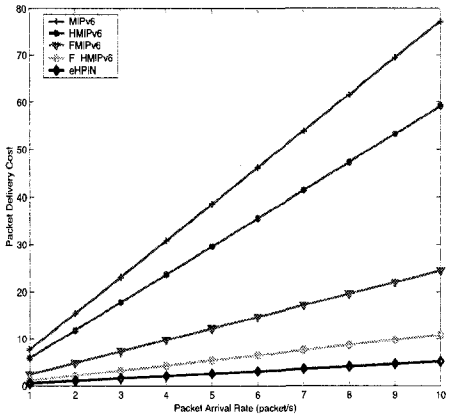

Figure 6.10 Impact of packet arrival rate on packet delivery cost.

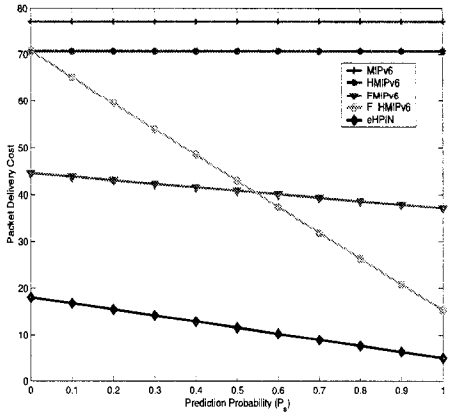

Figure 6.11 Impact of probability P_s on packet delivery cost.

Fig. 6.12 shows that the handoff latency increases proportionally with the wireless link delay. The handoff latency is very high for MIPv6 followed by HMIPv6 while FMIPv6 and F-HMIPv6 enable its reduction. eHPIN allows significant handoff latency reduction compared to other mobility management protocols. It is well known that the maximum tolerable delay for interactive conversation is approximately 200 ms. Hence, eHPIN can meet this requirement when the wireless link delay is set up below 50 ms. Fig. 6.13 shows the total packet loss in terms of packet arrival rate. Note that packet loss is much less prominent for eHPIN than for other IPv6-based handoff protocols. The effect of handoff in IPv6-based wireless environments is dominated by packet loss, which is due to the L2 handoff and the IP layer operations. In fact, due to the lack of any buffering and anticipated handoff

mechanisms in MIPv6 and HMIPv6, all in-flight packets are lost during handoff. However, in fast handoff schemes (i.e., FMIPv6, F-HMIPv6 and eHPIN) packet loss begins when L2 handoff is detected and until the buffering mechanism is initiated or if buffers overflow. Fig. 6.14 provides comparison of forwarded packets delay in-

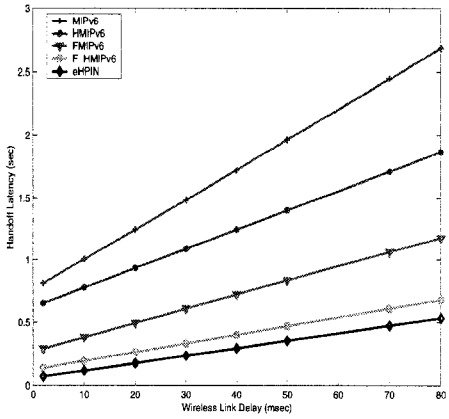

Figure 6.12 Handoff latency vs. wireless link delay.

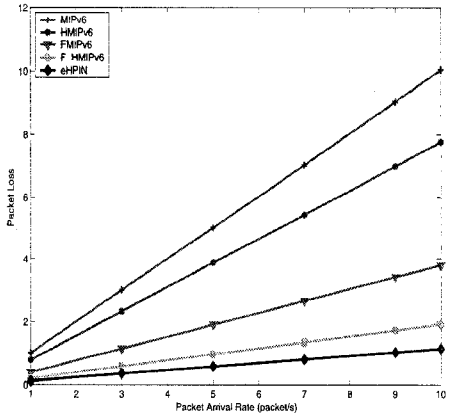

Figure 6.13 Packet loss vs. packet arrival rate.

duced by fast handoff schemes (i.e, FMIPv6, F-HMIPv6 and eHPIN). eHPIN allows the lowest delay for these packets, thus guaranteeing QoS for sessions with many forwarded packets. Fig. 6.15 shows that eHPIN has much lower handoff blocking probability than other IPv6-based handoff schemes. This result is due to the ability of eHPIN to reduce signal message exchanges and handoff latency. Thus, eHPIN can safely provide seamless handoff with service continuity.

6.7 Conclusion

The interworking of networks and mobility management are key issues in NGWN or 4G. Several proposals are available in the literature for these two issues. However, they fail to satisfy basic requirements such as seamless roaming and service continuity for real-time applications. This paper proposes an efficient handoff manage-

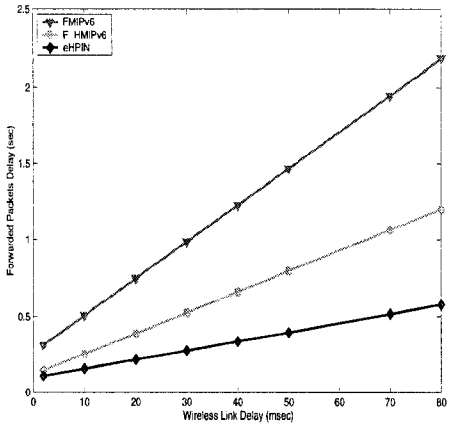

Figure 6.14 Forwarded packets delay vs. wireless link delay.

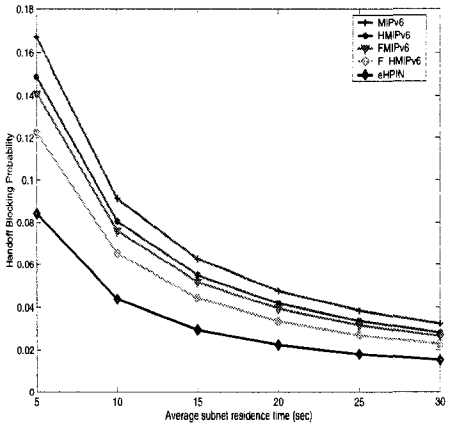

Figure 6.15 Comparison of handoff blocking probability.

ment protocol, called *enhanced Handoff Protocol for Integrated Networks* (eHPIN), to enable a better performance in IPv6-based heterogeneous wireless networks. eHPIN is a one-suite protocol to cope with access network discovery, fast handoff, context transfer and local mobility.

Performance analyses demonstrate a significant improvement for quality of service (QoS), which is defined in terms of signaling traffic overhead, packet delivery cost, handoff latency, packet loss and handoff blocking probability, compared with existing IPv6-based mobility management protocols. In other words, eHPIN alleviates services disruption and guarantees seamless roaming during handoff by allowing the selection of the best available network.

CHAPITRE 7

DISCUSSION GÉNÉRALE

Dans ce chapitre, nous commençons par une synthèse de nos objectifs de recherche et notre contribution au regard des différents défis évoqués dans la problématique. Par la suite, nous exposerons l'approche méthodologique considérée suivi de l'analyse des résultats obtenus dans leur ensemble. Enfin, nous terminerons par la portée de ces résultats.

7.1 Synthèse des travaux

La recherche menée dans cette thèse a donné lieu à quatre articles principaux de revues, un chapitre de livre et plusieurs articles de conférences internationales avec comité de lecture. Chacun de ces articles de revues traite un ou plusieurs points évoqués dans nos objectifs de recherche que nous allons récapituler dans les paragraphes ci-dessous. Deux des articles de revues ont déjà été acceptés pour publication tandis que les deux autres sont actuellement en cours d'évaluation.

Notre premier objectif de recherche portait sur l'analyse des mécanismes et protocoles de gestion de mobilité disponibles dans la littérature. Cet objectif a été réalisé grâce à une revue de littérature exhaustive sur les architectures d'intégration et les protocoles de gestion de mobilité pour les réseaux SFPG/4G. De plus, nous avons proposé un cadre ou modèle analytique robuste permettant d'évaluer les performances de ces mécanismes et protocoles.

Partant de l'analyse précédente, nos objectifs subséquents étaient de proposer

une architecture permettant une intégration transparente des réseaux d'accès distincts d'une part et de concevoir des mécanismes de gestion de mobilité efficaces et robustes d'autre part. La proposition d'une nouvelle architecture doit respecter certaines contraintes et exigences. Nous nous sommes donc basés sur cette logique afin de proposer une nouvelle architecture hybride assurant l'intégration des réseaux d'accès différents utilisants éventuellement des technologies distinctes et qui garantit leur interopérabilité. L'architecture proposée minimise autant que possible l'ajout de nouvelles entités, mais étend plutôt les fonctionnalités des entités existantes. En outre, elle permet de séparer le plan de contrôle (trafic de signalisation) du plan de transport (trafic des données). Ainsi, on a une architecture évolutive, fiable et économique.

Les protocoles et mécanismes de gestion de mobilité disponibles dans la littérature ne garantissent pas une itinérance sans coupure aux usagers. Nous avons donc traité cette question en proposant de nouveaux mécanismes et protocoles de gestion de mobilité. Dans les réseaux SFPG/4G, la décision de relève ne peut pas être basée uniquement sur la qualité de la puissance du signal ou la disponibilité de la bande passante comme c'est le cas dans les réseaux homogènes. Ainsi, nous avons proposé une nouvelle stratégie de décision de relève qui permet de prendre en compte plusieurs facteurs tels que le coût monétaire, la localisation géographique, le profil de l'usager en plus des deux facteurs précédemment cités.

Dans la littérature, le problème de gestion de mobilité est souvent traité sous différents aspects et indépendamment. En effet, on a par exemple un protocole pour chacun des problèmes suivants : découverte de réseau, transfert de contexte, anticipation de la relève et la mobilité locale. Notre approche consiste plutôt à proposer un protocole unifié qui permet de traiter conjointement toutes ces questions de manière intelligente et efficace. Les différentes versions des protocoles que nous avons proposées et validées permettent d'atteindre cet objectif en particulier dans le souci

d'offrir une meilleure QdS aux usagers et des meilleures performances réseau aux opérateurs. D'autre part, le protocole SIP est largement adopté pour supporter des services à valeur ajoutée. La combinaison de SIP et l'architecture IISA proposée permet de supporter et offrir encore plus de services et applications éventuelles.

Enfin, mentionnons que le draft portant sur le protocole *Proxy Mobile IPv6* ou PMIPv6 (Gundavelli *et al.* , 2007) a été publié presque à la fin de notre thèse. Plusieurs concepts que nous avons proposé dans les protocoles HPIN et eHPIN par exemple, la gestion des caches d'association par les routeurs d'accès ou *Access Edge Node* (AEN) sont aussi utilisés. De même, l'IDE (*Interworking Decision Engine*) permet d'effectuer une émulation du réseau nominal d'un MN qui est en déplacement dans un réseau visité ou étranger. Les mécanismes de gestion de relève actuellement déployer dans les réseaux 3GPP et 3GPP2 sont en général orienté réseau. Pour ce faire, les opérateurs des réseau 3GPP/3GPP2 et WiMAX manifestent un intérêt particulier pour une approche de gestion de mobilité orienté réseau au niveau de la couche IP. Les protocoles HPIN et eHPIN fournissent donc une solution pour un déploiement dans un contexte pratique.

7.2 Méthodologie

La proposition de nouveaux mécanismes doit être validée à l'aide d'une preuve de concepts. Nous avons choisi deux approches pour évaluer les performances des mécanismes et l'architecture que nous avons proposé. En effet, nous avons utilisé une modélisation analytique et une validation par simulation. Deux outils ont été utilisés pour y arriver à savoir, les logiciels MATLAB et OPNET. Tout d'abord, nous avons développé un modèle analytique comme fondement pour la première phase de validation. Ce modèle nous a permis d'étudier de manière approfondie les performances des protocoles existants dans la littérature ainsi que ceux que nous

avons proposé.

Le modèle est générique, il n'est donc pas limité à un protocole en particulier. Plusieurs batteries de tests ont été effectués pour la validation. L'absence de plusieurs modules dans le logiciel OPNET nous a contraint d'en ajouter de nouveaux et de les implémenter afin de pouvoir valider nos propositions. Cette tâche a été l'une des plus laborieuses durant notre recherche, car il a fallu ajouter plusieurs modules au préalable avant de commencer l'implémentation proprement dite de notre contribution. Toutefois, nous y sommes arrivés dans la mesure de notre possible.

7.3 Analyse des résultats

La validation numérique et par simulation des mécanismes, protocoles et architecture proposés montre qu'on obtient des résultats très satisfaisants. En effet, on a une architecture qui en plus de permettre l'intégration des réseaux, peut être déployée à des coûts moindres, ce qui est très avantageux pour les opérateurs dont le souci majeur est le retour sur l'investissement et la performance de leur réseau. L'architecture proposée permet de prendre en compte les opérateurs et fournisseurs déjà présents sur le marché ainsi que des nouveaux.

En effet, l'architecture proposée est évolutive, robuste et fiable. De plus, elle permet aux opérateurs et fournisseurs de services d'établir qu'une seule entente d'itinérance ou de service avec une tierce-partie (IDE) au lieu d'avoir des ententes directes avec plusieurs opérateurs. Le protocole eHPIN permet d'avoir une amériolation de performances comparativement à HPIN. En effet, on a une réduction du coût du trafic de signalisation d'environ 27%. Cette réduction est de 3% pour le délai de transfert des paquets tandis qu'elle est de 1% pour le coût de la livraison des paquets. Par contre, le délai de relève est presque similaire pour les deux pro-

tocoles. Les protocoles proposés permettent une réduction de la surcharge due à la tunnelisation sur la liaison sans fil.

Le mécanisme de décision de relève proposé permet d'obtenir une meilleure répartition de charge à travers les réseaux d'accès, ce qui induit des meilleurs débits pour les usagers. D'autre part, l'évaluation des performances montre qu'avec les protocoles et mécanismes proposés, le délai de relève, la perte des paquets, la probabilité de blocage des sessions et le trafic de signalisation sont réduits de façon considérable. On a ainsi une mobilité sans coupure et une continuité de services, en particulier pour des applications temps-réel. La comparaison avec les autres protocoles disponibles dans la littérature montre que notre proposition offre d'excellents résultats. Toutefois, pour certaines métriques, il y a des compromis à faire.

CHAPITRE 8

CONCLUSION ET RECOMMANDATIONS

Les réseaux de communication ne cessent d'évoluer et on s'oriente de plus en plus vers une convergence des réseaux fixe-mobile. Cette convergence entraîne le besoin d'intégration et d'interopérabilité entre les réseaux existants pour la définition et la conception des réseaux dits de prochaine ou quatrième génération. En outre, on constate un engouement vers des applications multimédia et des usagers qui deviennent très mobiles avec des exigences élevées sur la QdS auxquels ils ont souscrits. Il est donc crucial de résoudre toutes ces questions pour assurer le succès des réseaux SFPG/4G lors de leur déploiement. Cette thèse avait pour objectif d'apporter des solutions à ces différents problèmes en proposant une architecture intégrée et des mécanismes de gestion de mobilité dans un environnement sans fil et mobile hétérogène tout en offrant une QdS aux usagers. Dans ce chapitre qui nous permet de conclure nos travaux, nous allons mettre en évidence les contributions de cette thèse. Par la suite, nous exposerons les limites de notre travail avant de terminer par une ébauche de recommandations pour des travaux futurs.

8.1 Sommaire des contributions

Le but de cette thèse était de proposer des mécanismes efficaces de gestion de mobilité dans les réseaux SFPG/4G offrant des garanties de qualité de service (QdS) aux usagers ainsi que l'intégration et l'interopérabilité des systèmes de communication sans fil et mobile existants. Cet objectif a été atteint grâce à plusieurs contributions, lesquelles, à notre avis, serviront à la conception et au déploiement

des réseaux SFPG/4G. Les contributions essentielles de cette thèse peuvent être décrites comme suit :

- Proposition d'un cadre analytique pour évaluer les performances des protocoles de gestion de mobilité au niveau de la couche IP. Cette proposition est précédée d'une analyse rigoureuse des requis de performance qu'un protocole de gestion de mobilité devrait avoir et quels types de métriques il faut utiliser pour caractériser la QdS. L'interaction et l'influence de plusieurs facteurs sont prises en compte dans le cadre proposé afin de le rendre plus efficace et robuste comparativement aux études disponibles dans la littérature.
- Après l'analyse des architectures d'intégration disponibles et la caractérisation des exigences qu'une architecture devrait avoir, nous avons proposé une nouvelle architecture intégrée offrant une interopérabilité entre différents réseaux. Cette architecture satisfait les requis qui ont été identifiés. De plus, l'entité IDE introduite peut être intégrée comme fonctionnalité auprès d'un courtier (*Broker*) déjà existant.
- Proposition d'une fonction de décision de relève qui permet de prendre en compte plusieurs facteurs au niveau réseau et le profil des usagers pour la sélection du meilleur réseau d'accès. Cette fonction permet en outre une meilleure répartition de charges entre les réseaux d'accès.
- Conception d'une stratégie de gestion des interfaces radio afin de garantir un meilleur compromis entre la découverte de réseaux d'accès et la consommation d'énergie des terminaux mobile. Pour l'authentification des usagers en itinérance, un mécanisme basé sur l'utilisation d'un jeton (*token*) est proposé pour accélérer cette procédure.
- Proposition de trois versions de protocoles de gestion de mobilité au niveau IP. Ces protocoles utilisent les informations disponibles au niveau de la couche liaison pour assurer une anticipation de la relève afin de minimiser la latence et

la perte des paquets. L'évaluation d'autres métriques telles que le volume du trafic de signalisation et la probabilité d'échec de relève montre qu'on obtient des meilleurs résultats. Ces protocoles permettent d'assurer une mobilité sans coupure et une continuité de services dans les réseaux SFPG/4G.

- Conception d'un nouveau mécanisme pour la mise à jour des caches d'association et le transfert de contexte afin de réduire le trafic de signalisation sur la portion sans fil du réseau. Les routeurs d'accès et les routeurs passerelles sont utilisés comme agent proxy pour effectuer la mise à jour au nom des terminaux mobile.
- Toutes ces solutions ont été validées par simulation et analytiquement. Cette validation a permis d'avoir des résultats qui montrent une amélioration des performances par rapport aux protocoles et mécanismes disponibles dans la littérature. Les performances obtenues respectent les exigences et spécifications des applications tel que défini par les organismes de standardisation.

Enfin, comme nous l'avons déjà mentionné, cette thèse a donnée lieu à deux articles acceptés dans des revues/journaux, un chapitre de livre et cinq articles de conférences internationales avec comité de lecture. Quatre articles sont actuellement en cours d'évaluation, dont trois dans des revues et un pour une conférence. La liste de ces articles est donnée à la suite de ce chapitre.

8.2 Limitations des travaux

Les réseaux SFPG/4G étant encore dans une phase de conception, notre contribution ne peut prétendre avoir résolu tous les problèmes. Notre travail présente donc certaines limitations. Une première limitation pourrait venir de la réticence des opérateurs ou fournisseurs de services à collaborer avec une tierce-partie (IDE). Cependant, cette limitation peut être surmontée, car la plupart des grands opéra-

teurs (*majors* ou *seniors*) font partie des organismes de standardisation ou consortiums auxquels on pourrait attribuer la gestion de l'IDE.

La deuxième limitation qui ne dépend pas directement de notre travail est associée aux contraintes des logiciels de simulation offrant très peu de flexibilité. En effet, plusieurs modules ou protocoles ne sont pas disponibles. Il a fallu implémenter au préalable ces protocoles et mécanismes avant nos contributions pour effectuer les comparaisons. Cependant, avec des contraintes de temps, on ne pouvait pas implémenter de façon globale tous ces mécanismes, ce qui pourrait limiter l'évaluation des performances à grande échelle.

Une autre limitation serait associée à l'absence explicite d'un mécanisme pour la mise en correspondance des paramètres de QdS. En effet, nos solutions font cette hypothèse sans en donner la description. Cette correspondance est essentielle au vu de l'hétérogénéité des réseaux SFPG/4G. Enfin, une validation des différents mécanismes et protocoles dans un environnement réel aurait éventuellement permis d'avoir une idée par rapport aux résultats sur la fiabililté, l'évolutivité et les compromis de conception et d'implémentation. Notons toutefois qu'il n'est pas facile dans un environnement académique d'effectuer une expérimentation réelle sur des infrastructures réseaux. Les opérateurs ne sont pas prêts à divulguer certaines informations jugées confidentielles. D'où l'utilisation des simulations ou le prototypage.

8.3 Indication des travaux futurs

Il reste encore beaucoup de choses à faire pour la conception des réseaux SFPG ou 4G. Nous présentons ci-après quelques pistes de travaux futurs qui pourraient s'inscrire dans la continuité logique de cette thèse.

Une possibilité serait d'incorporer un mécanisme de réservation de ressources au niveau du réseau d'accès et une stratégie de différentiation de services dans le réseau cœur dans le souci d'offrir une QdS adéquate de bout-en-bout. Les protocoles proposés assumaient une interaction des paramètres de QdS entre les différents réseaux d'accès. Il serait intéressant de définir un mécanisme permettant la mise en correspondance des paramètres de QdS entre les différentes technologies des réseaux d'accès. D'autre part, le nouvel élément introduit, *Interworking Decision Engine* (IDE), peut être une entité logique ou physique. Dans le dernier cas, une étude de son emplacement à travers le réseau pourrait s'avérer intéressante.

Le contrôle d'admission devrait aussi être examiné comme politique de gestion des ressources pour une étude plus approfondie de la probabilité de blocage des sessions ou de leur interruption forcée. Notre étude portait essentiellement sur l'interaction entre les couches liaison de données et IP pour la gestion de mobilité. Cependant, cette interaction pourrait s'étendre aux couches de niveaux supérieurs tels que transport et application pour décider du meilleur instant de relève afin d'assurer une gestion de mobilité plus fiable et stable.

La fonction de décision de relève proposée requiert l'information sur l'importance relative de chaque facteur. Cette importance est donnée par des pondérations (poids). Nous avons considéré un ensemble de poids statiques. Ces poids ayant un impact sur la performance de la décision de relève, il serait intéressant de proposer des techniques permettant de déterminer ou sélectionner des valeurs optimales pour lesdits poids. Enfin, sur un plan plus pratique ou opérationnel, il serait souhaitable d'effectuer des tests plus approfondis sur des environnements réels, par exemple via un émulateur de réseaux ou par prototypage.

DIFFUSION DES RÉSULTATS

- **Chapitre de livre :**

  L1)- **C. Makaya** and S. Pierre : "Trends and Challenges for Mobility Management in IP-based Next Generation Wireless Networks," in *Wireless Communications Research Trends*, Tong S. Lee (Ed.), Nova Science Publisher, New York, USA, à paraître en 2007 (3rd Quarter), ISBN : 1-60021-674-9.

- **Articles de revues/journaux :**

  J1)- **C. Makaya** and S. Pierre : "Integrated Architecture for Heterogeneous Wireless Networks," article invité, *Journal of Networks*, Juillet 2007.

  J2)- **C. Makaya** and S. Pierre : "Enhanced Fast Handoff Scheme for IP-based Next Generation Wireless Networks," soumis pour évaluation, *Computer Communications (Elsevier)*, Avril 2007.

  J3)- **C. Makaya** and S. Pierre : "Adaptive Handoff Scheme for Heterogeneous IP Wireless Networks," soumis pour évaluation, *Computer Communications (Elsevier)*, Avril 2007.

  J4)- **C. Makaya** and S. Pierre : "An Architecture for Seamless Mobility Support in IP-based Next-Generation Wireless Networks," *IEEE Transactions on Vehicular Technology*, accepté pour publication en Juin 2007.

  J5)- **C. Makaya** and S. Pierre : "An Analytical Framework for Performance Evaluation of IPv6-based Mobility Management Protocols," *IEEE Transactions on Wireless Communications*, accepté pour publication en Mars 2007.

- **Articles de conférences :**

  C1)- **C. Makaya** and S. Pierre : "Dynamic Vertical Handoff Protocol for

Heterogeneous Wireless Networks," soumis pour évaluation, Juin 2007.

C2)- **C. Makaya** and S. Pierre : "IP-based Fast Handoff Scheme for Heterogeneous Wireless Networks," *Proc. of IEEE International Conference on Wireless and Mobile Computing, Networking and Communications* (WiMob'07), White Plains, New York, USA, October 8-10, 2007.

C3)- **C. Makaya** and S. Pierre : "Handoff Performance Enhancement in Wireless Overlay Networks," *Proc. of IASTED International Conference on Wireless and Optical Communications* (WOC'07), pp. 37-42, Montreal, Quebec, Canada, May 30-June 1, 2007.

C4)- **C. Makaya** and S. Pierre : "Efficient Handoff Scheme for Heterogeneous IPv6-based Wireless Networks," *Proc. of IEEE Wireless Communications and Networking Conference* (WCNC'07), pp. 3258-3263, Hong Kong, China, March 11-15, 2007.

C5)- **C. Makaya** and S. Pierre : "Interworking Architecture for Heterogeneous IP Wireless Networks," *Proc. of IARIA International Conference on Wireless and Mobile Communications* (ICWMC'07), Guadeloupe, French Caribbean, March 4-9, 2007.

C6)- **C. Makaya** and S. Pierre : "Handoff Protocol for Heterogeneous All-IP Wireless Networks," *Proc. of IEEE Canadian Conference on Electrical and Computer Engineering* (CCECE'06), pp. 223-226, Ottawa, Ontario, Canada, May 7-10, 2006.

RÉFÉRENCES

3GPP. 2003. *Feasability Study on 3GPP System to WLAN Interworking (Release 6)*. 3GPP TR 22.934 v6.2.0. 3rd Generation Partnership Project (3GPP). 99p.

3GPP. 2004. *3GPP System to WLAN Interworking; System Description (Release 6)*. 3GPP TS 23.234 v6.3.0. 3rd Generation Partnership Project (3GPP). 30p.

3GPP2. 2004. *3GPP2-WLAN Interworking; Stage 1 Requirements*. 3GPP2 TS S.R0087-0 v1.0. 3rd Generation Partnership Project 2 (3GPP2). 19p.

3GPP2. 2006. *cdma2000-WLAN Interworking; Stage 1 Requirements*. 3GPP2 TS S.R0087-A v1.0. 3rd Generation Partnership Project 2 (3GPP2). 22p.

ABOBA, B., & BEADLES, M. 1999. *The Network Access Identifier*. RFC 2486. Internet Engineering Task Force (IETF). 8p.

AHMAVAARA, K., HAVERINEN, H., & PICHNA, R. 2003. "Interworking Architecture Between 3GPP and WLAN Systems". *IEEE Communications Magazine*, **41**(11), 74–81.

AKYILDIZ, I. F., MCNAIR, J., HO, J. S. M., UZUNALIOGLU, H., & WANG, W. 1999. "Mobility Management in Next Generation Wireless Systems". *Proc. of IEEE*, **87**(8), 1347–1384.

AKYILDIZ, I. F., XIE, J., & MOHANTY, S. 2004. "A Survey of Mobility Management in Next-Generation All-IP-Based Wireless Systems". *IEEE Wireless Communications*, **11**(4), 16–28.

AKYILDIZ, I. F., MOHANTY, S., & XIE, J. 2005. "A Ubiquitous Mobile Communication Architecture for Next-Generation Heterogeneous Wireless Systems". *IEEE Communigations Magazine*, **43**(6), S29–S36.

AL-GIZAWI, T., AXIOTIS, D. I., PEPPAS, K., LAZARAKIS, F., & VLAHODIMITROPOULOS, K. 2002. "Definition of 4G scenarios, traffic and mobility characteristics for interoperating UMTS and WLAN networks". *7th Wireless World Research Forum Meeting (WWRF'02)*, 1–9.

ASSOUMA, A. D., BEAUBRUN, R., & PIERRE, S. 2006. "Mobility Management in Heterogeneous Wireless Networks". *IEEE Journal on Selected Areas in Communications*, **24**(3), 638–648.

BAUMANN, F. V., & NIEMEGEERS, I. G. 1994. "An Evaluation of Location Management Procedures". *Proc. of 3rd Annual International Conference Universal Personal Communications (UPC'94)*, 359–364.

BEAUBRUN, R., PIERRE, S., & CONAN, J. 2005. "An Approach for Managing Global Mobility and Roaming in the Next-Generation Wireless System". *Computer Communications*, **28**(5), 571–581.

BUDDHIKOT, M., CHANDRANMENON, G., HAN, S., LEE, Y. W., MILLER, S., & SALGARELLI, L. 2003. "Integration of 802.11 and Third-Generation Wireless Data Networks". *Proc. of IEEE International Conference on Computer Communications (INFOCOM'03)*, **1**, 503–512.

CAMPBELL, A. T., GOMEZ, J., SANGHYO, K., CHIEH-YIH, W., TURANYI, Z. R., & VALKO, A. G. 2002. "Comparison of IP Micromobility Protocols". *IEEE Wireless Communications*, **9**(1), 72–82.

CASTELLUCCIA, C. 2000. "HMIPv6 : A Hierarchical Mobile IPv6 Proposal". *ACM SIGMOBILE Mobile Computing and Communications Review*, **4**(1), 48–59.

CHEN, J.-C., & ZHANG, T. 2004. *IP-Based Next-Generation Wireless Networks : Systems, Architectures and Protocols.* Hoboken, New Jersey, USA : John Wiley & Sons, Inc.

CHIUSSI, F. M., KHOTIMSKY, D. A., & KRISHNAN, S. 2002. "Mobility Management in Third-Generation All-IP Networks". *IEEE Communications Magazine*, **40**(9), 124–135.

DAS, S., MCAULEY, A., A., DUTTA, MISRA, A., CHAKRABORTY, K., & DAS, S. K. 2002. "IDMP : An Intradomain Mobility Management Protocol for Next-Generation Wireless Networks". *IEEE Wireless Communications*, **9**(3), 38–45.

DEERING, S., & HINDEN, R. 1998. *Internet Protocol, version 6 (IPv6) Specification.* RFC 2460. Internet Engineering Task Force (IETF). 39p.

DU, F., NI, L. M., & ESFAHANIAN, A. H. 2002. "HOPOVER : A new Handoff Protocol for Overlay Networks". *Proc. of IEEE International Conference on Communications (ICC'02)*, **5**, 3234–3239.

FANG, Y. 2003. "Movement-Based Mobility Management and Trade Off Analysis for Wireless Mobile Networks". *IEEE Transactions on Computers*, **52**(6), 791–803.

GUNDAVELLI, S., LEUNG, K., DEVARAPALI, V., CHOWDHURY, K., & PATIL, B. 2007. *Proxy Mobile IPv6.* Internet Draft draft-ietf-netlmm-proxymip6-06.txt. Internet Engineering Task Force (IETF). 58p. work in progress.

GUSTAFSSON, E., & JONSSON, A. 2003. "Always Best Connected". *IEEE Wireless Communications*, **10**(1), 49–55.

GWON, Y., KEMPF, J., & YEGIN, A. 2004. "Scalability and Robustness Analysis of Mobile IPv6, Fast Mobile IPv6, Hierarchical Mobile IPv6, and Hybrid IPv6 Mobility Protocols using a Large-scale Simulation". *Proc. of IEEE International Conference on Communications (ICC'04)*, **7**, 4087–4091.

HEDRICK, C. 1988. *Routing Information Protocol.* RFC 1058. Internet Engineering Task Force (IETF). 33p.

HSIEH, R., ZHOU, Z. G., & SENEVIRATNE, A. 2003. "S-MIP : A Seamless Handoff Architecture for Mobile IP". *Proc. of IEEE International Conference on Computer Communications (INFOCOM'03)*, **3**, 1774–1784.

HUI, S. Y., & YEUNG, K. H. 2003. "Challenges in the Migration to 4G Mobile Systems". *IEEE Communications Magazine*, **41**(12), 54–59.

JASEEMUDDIN, M. 2003. "An Architecture for Integrating UMTS and 802.11 WLAN Networks". *Proc. of IEEE Symposium on Computers and Communications (ISCC'03)*, 716–723.

JOHNSON, D. B., PERKINS, C. E., & ARKKO, J. 2004. *Mobility Support in IPv6.* RFC 3775. Internet Engineering Task Force (IETF). 165p.

JUNG, H. Y., KIM, E. A., YI, J. W., & LEE, H. H. 2005a. "A Scheme for Supporting Fast Handover in Hierarchical Mobile IPv6 Networks". *ETRI Journal*, **27**(6), 798–801.

JUNG, H. Y., SOLIMAN, H., KOH, S. J., & TAKAMIYA, N. 2005b. *Fast Handover for Hierarchical MIPv6 (F-HMIPv6).* Internet Draft draft-jung-mipshop-fhmipv6-00.txt. Internet Engineering Task Force (IETF). 18p. work in progress.

KEMPF, J., WOOD, J., & FU, G. 2003. "Fast Mobile IPv6 Handover Packet Loss Performance : measurement for emulated real time traffic". *Proc. of IEEE Wireless Communications and Networking Conference (WCNC'03)*, **2**, 1230–1235.

KEMPF, J. (ED.). 2007a. *Goals for Network-Based Localized Mobility Management (NETLMM).* RFC 4831. Internet Engineering Task Force (IETF). 14p.

KEMPF, J. (ED.). 2007b. *Problem Statement for Network-Based Localized Mobility Management (NETLMM).* RFC 4830. Internet Engineering Task Force (IETF). 13p.

KENT, S., & ATKINSON, R. 1998. *Security Architecture for the Internet Protocol.* IETF RFC 2401. Internet Engineering Task Force (IETF). 66p.

KIBRIA, M. R., & JAMALIPOUR, A. 2007. "On Designing Issues of the Next Generation Mobile Network". *IEEE Network*, **21**(1), 6–13.

KOH, S. J., CHANG, M. J., & LEE, M. 2004. "mSCTP for Soft Handover in Transport Layer". *IEEE Communications Letters*, **8**(3), 189–191.

KOH, S. J., XIE, Q., & PARK, S. D. 2005. *Mobile SCTP (mSCTP) for IP Handoff Support.* IETF Draft draft-sjkoh-msctp-01.txt. Internet Engineering Task Force (IETF). 19p. work in progress.

KOODLI, R., & PERKINS, C. E. 2001. "Fast Handovers and Context Transfers in Mobile Networks". *ACM SIGCOMM Computer Communication Review*, **31**(5), 37–47.

KOODLI, R. (ED.). 2005. *Fast Handovers for Mobile IPv6.* RFC 4068. Internet Engineering Task Force (IETF). 42p.

LAI, W. K., & CHIU, J. C. 2005. "Improving Handoff Performance in Wireless Overlay Networks by Switching Between Two-Layer IPv6 and One-Layer IPv6 Addressing". *IEEE Journal on Selected Areas in Communications*, **23**(11), 2129–2137.

LAMPROPOULOS, G., PASSAS, N., L., MERAKOS, & KALOXYLOS, A. 2005. "Handover Management Architecture in Integrated WLAN/Cellular Networks". *IEEE Communications Surveys & Tutorials*, **7**(4), 30–44.

LEIBSCH, M., SINGH, A., CHASKAR, H., FUNATO, D., & SHIM, E. 2005. *Candidate Access Router Discovery (CARD).* RFC 4066. Internet Engineering Task Force (IETF). 46p.

LOUGHNEY, J., NAKHJIRI, M., PERKINS, C., & KOODLI, R. 2005. *Context Transfer Protocol (CXTP).* RFC 4067. Internet Engineering Task Force (IETF). 33p.

MAKAYA, C., & PIERRE, S. 2006. "Handoff Protocol for Heterogeneous All-IP-based Wireless Networks". *Proc. of IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'06)*, 223–226.

MAKAYA, C., & PIERRE, S. 2007a. "An Analytical Framework for Performance Evaluation of IPv6-based Mobility Management Protocols". *IEEE Transactions on Wireless Communications*, to appear.

MAKAYA, C., & PIERRE, S. 2007b. "An Interworking Architecture for Heterogeneous IP Wireless Networks". *Proc. of IARIA International Conference on Wireless and Mobile Communications (ICWMC'07).*

MCNAIR, J., & ZHU, F. 2004. "Vertical Handoffs in Fourth-Generation Multinetworks Environments". *IEEE Wireless Communications*, **11**(3), 8–15.

MCNAIR, J., AKYILDIZ, I. F., & BENDER, M. D. 2001. "Handoffs for Real-Time Traffic in Mobile IP version 6 Networks". *Proc. of IEEE Global Telecommunication Conference (GLOBECOM'01)*, **6**, 3463–3467.

MINJI, N., NAKJUNG, C., YONGHO, S., & YANGHEE, C. 2004. "WISE : Energy-efficient Interface Selection on Vertical Handoff between 3G Networks and WLANs". *Proc. of IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC'04)*, **1**, 692–698.

MISRA, A., DAS, S., A., DUTTA, MCAULEY, A., & DAS, S. K. 2002. "IDMP-based Fast Handoffs and Paging in IP-based 4G Mobile Networks". *IEEE Communications Magazine*, **40**(3), 138–145.

MOORE, N. 2006. *Optimistic Duplicate Address Detection.* RFC 4429. Internet Engineering Task Force (IETF). 17p.

NASSER, N., HASSWA, A., & H., HASSANEIN. 2006. "Handoffs in Fourth Generation Heterogeneous Networks". *IEEE Communications Magazine*, **44**(10), 96–103.

ONG, L., & YOAKUM, J. 2002. *An introduction to the Stream Control Transmission Protocol (SCTP).* RFC 3286. Internet Engineering Task Force (IETF). 10p.

PACK, S., & CHOI, Y. 2003. "Performance Analysis of Fast Handover in Mobile IPv6 Networks". *Proc. of IFIP Personal Wireless Communications, Lecture Notes in Computer Science (LNCS)*, **2775**, 679–691.

PACK, S., & CHOI, Y. 2004. "A Study on Performance of Hierarchical Mobile IPv6 in IP-based Cellular Networks". *IEICE Transactions on Communications*, **E87-B**(3), 462–469.

PÉREZ-COSTA, X., SCHMITZ, R., HARTENSTEIN, H., & LEIBSCH, M. 2002. "MIPv6, FMIPv6 and HMIPv6 Handover Latency Study : Analytical Approach". *Proc. of the IST Mobile and Wireless Communications Summit*, 100–105.

PÉREZ-COSTA, X., TORRENT-MORENO, M., & HARTENSTEIN, H. 2003. "A Performance Comparison of Mobile IPv6, Hierarchical Mobile IPv6, Fast Handovers for Mobile IPv6 and their Combination". *ACM SIGMOBILE Mobile Computing and Communications Review*, **7**(4), 5–19.

REINBOLD, P., & BONAVENTURE, O. 2003. "IP Micro-Mobility Protocols". *IEEE Comunications Surveys & Tutorials*, **5**(1), 40–57.

RIEGEL, M., & TUEXEN, M. 2006. *Mobile SCTP.* Internet Draft draft-riegel-tuexen-mobile-sctp-07.txt. Internet Engineering Task Force (IETF). 13p. work in progress.

ROBERTS, M. L., TEMPLE, M. A., MILLS, R. F., & RAINES, R. A. 2006. "Evolution of the Air Interface of Cellular Communications Systems Toward 4G Realization". *IEEE Communications Surveys & Tutorials*, **8**(1), 2–23.

ROSENBERG, J., SCHULZRINNE, H., CAMARILLO, G., JOHNSTON, A., PETERSON, J., SPARKS, R., HANDLEY, M., & SCHOOLER, E. 2002. *SIP : Session Initiation Protocol.* RFC 3261. Internet Engineering Task Force (IETF). 202p.

SAHA, D., MUKHERJEE, A., MISRA, I. S., CHAKRABORTY, M., & SUBHSAH, N. 2004. "Mobility Support in IP : a Survey of Related Protocols". *IEEE Network*, **18**(6), 34–40.

SALKINTZIS, A. K., FORS, C., & PAZHYANNUR, R. 2002. "WLAN-GPRS Integration for 3G Next-Generation Mobile Data Networks". *IEEE Wireless Communications*, **9**(5), 112–124.

SCHULZRINNE, H., & WEDLUND, E. 2000. "Application-layer Mobility using SIP". *ACM SIGMOBILE Mobile Computing and Communications Review*, **4**(3), 47–57.

SHENOY, N. 2005. "A Framework for Seamless Roaming Across Heterogeneous Next Generation Wireless Networks". *Wireless Networks*, **11**(6), 757–774.

SOLIMAN, H., CASTELLUCCIA, C., EL-MALKI, K., & BELLIER, L. 2005. *Hierarchical Mobile IPv6 Mobility Management (HMIPv6).* RFC 4140. Internet Engineering Task Force (IETF). 29p.

SONG, J. Y., LEE, S. W., & CHO, D. H. 2003. "Hybrid Coupling Scheme for UMTS and Wireless LAN Interworkings". *Proc. of IEEE Vehicular Technology Conference (VTC'03)-Fall*, **4**, 2247–2251.

SONG, Q., & JAMALIPOUR, A. 2005. "Network Selection in an Integrated Wireless LAN and UMTS Environment using Mathematical Modeling and Computing Techniques". *IEEE Wireless Communications*, **12**(3), 42–48.

STEMM, M., & KATZ, R. H. 1998. "Vertical Handoffs in Wireless Overlay Networks". *ACM Mobile Networking and Applications (MONET)*, **3**(4), 335–350.

TAFAZOLLI, R., MOESSNER, K., POLITIS, C., & DAGIUKLAS, T. 2005. Cooperation Between Networks. *Chap. 4, pages 175–248 of :* TAFAZOLLI, R. (ed), *Technologies for the Wireless Future.* West Sussex, England : John Wiley & Sons, Ltd.

THOMSON, S., & NARTEN, T. 1998. *IPv6 Stateless Address Autoconfiguration.* RFC 2462. Internet Engineering Task Force (IETF). 25p.

VIVALDI, I., ALI, B. M., PRAKASH, V., & SALI, A. 2003. "Routing Scheme for Macro Mobility Handover in Hierarchical Mobile IPv6 Network". *Proc. of 4th National Conference on Telecommunication Technology (NCTT'03)*, 88–92.

WANG, H. J., KATZ, R. H., & GIESE, J. 1999. "Policy-Enabled Handoffs Across Heterogeneous Wireless Networks". *Proc. of the 2nd IEEE Workshop on Mobile Computing Systems and Applications (WMCSA'99)*, 51–60.

WANG, W., & AKYILDIZ, I. F. 2000. "Intersystem Location Update and Paging Schemes for Multitier Wireless Networks". *Proc. of ACM International Conference on Mobile Computing and Networking (MOBICOM'00)*, 99–109.

WANG, W., & AKYILDIZ, I. F. 2001. "A New Signaling Protocol for Intersystem Roaming in Next-Generation Wireless Systems". *IEEE Journal on Selected Areas in Communications*, **19**(10), 2040–2052.

WONG, K. D., DUTTA, A., SCHULZRINNE, H., & YOUNG, K. 2007. "Simultaneous Mobility : Analytical Framework, Theorems and Solutions". *Wireless Communications and Mobile Computing (Wiley)*, **7**(5), 623–642.

XIAO, Y., PAN, Y., & LIE, J. 2004. "Design and Analysis of Location Management for 3G Cellular Networks". *IEEE Transactions on Parallel and Distributed Systems*, **15**(4), 339–349.

XIE, J., & AKYILDIZ, I. F. 2002. "A Novel Distributed Dynamic Location Management Scheme for Minimizing Signaling Costs in Mobile IP". *IEEE Transactions on Mobile Computing*, **1**(3), 163–175.

YAGER, R. R. 1988. "On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making". *IEEE Transactions on Systems, Man and Cybernetics*, **18**(1), 183–190.

YAVATKAR, R., PENDARAKIS, D., & GUERIN, R. 2000. *A Framework for Policy-based Admission Control.* RFC 2753. Internet Engineering Task Force (IETF). 20p.

ZHU, F., & MCNAIR, J. 2006. "Multiservice Vertical Handoff Decision Algorithms". *EURASIP Journal on Wireless Communications and Networking*, **2006**.